ΜΕΘΟΔΟΙ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΓΙΑ ΤΟΝ ΕΝΤΟΠΙΣΜΟ ΜΟΤΙΒΩΝ ΣΕ ΣΥΜΒΟΛΟΣΕΙΡΕΣ

Η

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ ΕΞΕΙΔΙΚΕΥΣΗΣ

Υποβάλλεται στην

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

από την

Έλλη Βουδιγάρη

ως μέρος των Υποχρεώσεων
για τη λήψη
του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

ΜΕ ΕΞΕΙΔΙΚΕΥΣΗ ΣΤΙΣ ΤΕΧΝΟΛΟΓΙΕΣ-ΕΦΑΡΜΟΓΕΣ

Οκτώβριος 2007

ΑΦΙΕΡΩΣΗ

Στους αγαπημένους μου γονείς,

Νίκο

και

Δήμητρα

ΕΥΧΑΡΙΣΤΙΕΣ

Η διατριβή αυτή εκπονήθηκε στο τμήμα Πληροφορικής του Πανεπιστημίου Ιωαννίνων με επιβλέποντα τον Λέκτορα κ. Κωνσταντίνο Μπλέκα.

Θα ήθελα, καταρχήν, να ευχαριστήσω θερμά τον κ. Κωνσταντίνο Μπλέκα, επιβλέποντα της διατριβής μου και καθηγητή μου από το πρώτο εξάμηνο εισαγωγής μου στο μεταπτυχιακό πρόγραμμα σπουδών του τμήματος Πληροφορικής, για την άριστη επιλογή του θέματος, το οποίο κατάφερε να κρατήσει άγρυπνο το ενδιαφέρον μου καθ' όλη τη διάρκεια ενασχόλησής μου με την μεταπτυχιακή μου εργασία, τη συνεχή και άψογη καθοδήγησή του, καθώς και την εμπιστοσύνη που μου έδειξε, όσον αφορά το ερευνητικό κομμάτι της διατριβής.

Ακόμα, αισθάνομαι την ανάγκη να ευχαριστήσω τους γονείς μου, γιατί παρά το γεγονός ότι βρίσκονταν μακριά μου τα χρόνια των μεταπτυχιακών μου σπουδών, ήταν συγχρόνως και τόσο κοντά μου με την αγάπη και την υποστήριξη που μου προσέφεραν.

Τέλος, θα ήθελα να ευχαριστήσω την συμφοιτήτρια και πολύ αγαπημένη μου φίλη Μαρία Χριστοδουλίδου για την ειλικρινή υποστήριξη και συμπαράστασή της κάθε στιγμή που τη χρειαζόμουν, καθώς και τις, επί πολλών ετών φίλες μου και τέως συμφοιτήτριες από το Μαθηματικό τμήμα, Αγγελική Βιδάλη, Μαρία Πεντάρη, Δήμητρα Ζαφειροπούλου και Νέλη Αργυροπούλου, οι οποίες αν και μακριά με ενθάρρυναν και μου συμπαραστάθηκαν από την αρχή έως το πέρας των μεταπτυχιακών μου σπουδών.

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ

ΕΥΡΕΤΗΡΙΟ ΣΧΗΜΑΤΩΝ

ΠΕΡΙΛΗΨΗ

Έλλη Βουδιγάρη του Νικολάου και της Δήμητρας.
MSc, Τμήμα Πληροφορικής, Πανεπιστήμιο Ιωαννίνων, Οκτώβριος, 2007.
Τίτλος: Μέθοδοι Μηχανικής Μάθησης για τον Εντοπισμό Μοτίβων σε Συμβολοσειρές.
Επιβλέπων: Κωνσταντίνος Μπλέκας

Η παρούσα διατριβή πραγματεύεται το πρόβλημα της ανίχνευσης μιας ακολουθίας συμβόλων καθορισμένου μήκους (*μοτίβου*) που επαναλαμβάνεται στατιστικά μέσα σε ένα σύνολο συμβολοσειρών μεταβλητού μήκους, αποτελούμενων από σύμβολα ενός διακριτού αλφάβητου. Αρχικά, αναπτύξαμε γνωστές μεθόδους Μηχανικής Μάθησης: μικτά μοντέλα εκπαιδευόμενα με τον αλγόριθμο ΕΜ και δειγματολήπτη Gibbs (Gibbs Sampling), όπως επίσης και τον αλγόριθμο Build-PST που προσφέρει μία διαφορετική προσέγγιση μέσω της κατασκευής ενός πιθανοτικού δέντρου προθεμάτων.

Στη διατριβή, προτείνονται δύο νέες μέθοδοι που αποτελούν επεκτάσεις του βασικού σχήματος των μικτών μοντέλων με τον αλγόριθμο ΕΜ. Συγκεκριμένα, οι τεχνικές αυτές βασίζονται στην επέκταση (*μέθοδος του εκτεταμένου πίνακα*) ή οριοθέτηση (*μέθοδος του οριοθετούμενου πίνακα*) του στοχαστικού πίνακα περιγραφής του μοτίβου, προσφέροντας έτσι (έμμεσα) χωρικές ιδιότητες (τελεστές) στο στοχαστικό μοντέλο. Η πειραματική μελέτη έδειξε ότι, σε πολλές περιπτώσεις, οι προτεινόμενες τεχνικές εξασφαλίζουν βελτιωμένες λύσεις με βάση ποιοτικά και ποσοτικά κριτήρια και ελαττώνουν σε μεγάλο βαθμό την επίδραση της αρχικοποίησης του αλγορίθμου ΕΜ.

EXTENDED ABSTRACT IN ENGLISH

Voudigari Elli, N.
Msc, Computer Science Department, University of Ioannina, Greece. October, 2007.
Title: Machine Learning Methods for Pattern Discovering in Sequences.
Supervisor: Constantinos Blekas.

In this dissertation, we studied the problem of discovering the most common repeated pattern of fixed length in a set of variable length sequences consisted by symbols of a discrete alphabet. At first, we gave a general description of the problem and explained how a motif is described by a position weight matrix (pwm). The next chapters include the presentation of well-known and lately proposed methods used to determine approximate solutions of the examined problems, an analytical description of experiments realized in order to compare them and the conclusions that came out.

In Chapter 2, we presented two well-known machine learning methods usually used to determine an approximate solution in such kind of problems: EM and Gibbs Sampling. Specifically, this can be done through the training of a mixture model. We also described a different approach lately proposed, which offers satisfying results in the same research area. The last one is called as the Build – PST algorithm and tries to approximate the motif searched through the construction of a probability suffix tree. Besides the general description of the above methods, we explained in details how they are adapted in the examined problem.

Chapter 3 includes the general description of two well-known algorithms used for data clustering, K-means (partitional clustering) and Agglomerative (hierarchical clustering), and

the exact way we used them in order to take an initial approximation of the parameter of interest (i.e. the pwm describing the motif).

The most important part of the dissertation is consisted by the presentation of two new methods proposed in order to solve the same problem (Chapter4). These came out as extensions of the EM algorithm and take into account space information that the same method discards (disadvantages of EM).

In Chapter 5, the above methods are tested through the use of artificial and real data sets of different kind of motifs and different type of algorithms determining the initial parameter (Agglomerative and K-means), and we present the interesting results that came out.

In Chapter 6, we explain how the experiments realized offer a new perspective for future work in the same research area.

ΚΕΦΑΛΑΙΟ 1. ΕΙΣΑΓΩΓΗ

Ο εντοπισμός *μοτίβων*, δηλαδή επαναλαμβανόμενων ακολουθιών από γράμματα (ή σύμβολα), μέσα σε ένα σύνολο συμβολοσειρών μεγάλου και μεταβλητού μήκους, αποτελεί ένα σημαντικό πρόβλημα, για την επίλυση του οποίου έχουν αναπτυχθεί μια σειρά από αλγοριθμικές μέθοδοι τα τελευταία χρόνια. Καθοριστικό κίνητρο προς την κατεύθυνση αυτή, υπήρξε το γεγονός ότι το συγκεκριμένο πρόβλημα παρουσιάζει εφαρμογές σε μια πληθώρα ερευνητικών περιοχών, με κυριότερο τον συνεχώς εξελισσόμενο κλάδο της Βιολογίας. Συγκεκριμένα, η λύση αυτού του προβλήματος ενδέχεται να συμβάλει στην ανακάλυψη επαναλαμβανόμενων μοτίβων, πρωτεύουσας σημασίας (π.χ. σε πρωτεϊνικές αλυσίδες και σε αλυσίδες DNA), τα οποία μπορεί κάλλιστα να αποτελούν το αίτιο ύπαρξης ανίατων μέχρι στιγμής ασθενειών και να συντελέσει στην αποτελεσματική θεραπεία τους. Στη γενικότερη, τώρα, περίπτωση, ένα μοτίβο μπορεί να εσωκλείει μια καθοριστικής σημασίας πληροφορία, καθώς είναι πολύ πιθανό να χαρακτηρίζει μονοσήμαντα ένα σύνολο αλληλουχιών. Συνεπώς, η εξόρυξή του ενδέχεται να συμβάλει στην πληρέστερη μελέτη και στην παραγωγή νέας γνώσης μιας ολόκληρης οικογένειας ακολουθιών.

Πριν περιγράψουμε το κυρίως πρόβλημα που θα μελετήσουμε, αξίζει να κάνουμε μια σύντομη αναφορά στις κατηγορίες των μοτίβων που υπάρχουν, καθώς και να εισαγάγουμε κάποιες έννοιες, οι οποίες θα διευκολύνουν τον αναγνώστη στην πληρέστερη κατανόηση όσων ακολουθούν.

Καταρχήν, μπορούμε να διακρίνουμε δύο είδη μοτίβων, τα ντετερμινιστικά και τα πιθανοτικά [8]. Με τον όρο *ντετερμινιστικό* μοτίβο εννοούμε μια συγκεκριμένη ακολουθία χαρακτήρων, δηλαδή δοθείσης μιας συμβολοσειράς (ίδιου μήκους με το μοτίβο), αυτή είτε ανήκει σε αυτή την κατηγορία (εάν συμπίπτει με το μοτίβο) είτε όχι. Από την άλλη πλευρά, τα *πιθανοτικά* μοτίβα περιγράφονται στοχαστικά με μια κατανομή, με βάση την οποία προσδίδουν σε κάθε ακολουθία χαρακτήρων μια πιθανότητα να περιέχει ή όχι το μοτίβο. Ασφαλώς, σε αυτή την περίπτωση, ενδιαφερόμαστε για τον εντοπισμό των ακολουθιών, για τις οποίες αυτή η πιθανότητα γίνεται μέγιστη. Επιπλέον, είναι δυνατόν να υπάρχει κενό (gap) σε ορισμένες θέσεις ενός μοτίβου, γεγονός που σημαίνει ότι σε αυτή τη θέση μπορεί να αντιστοιχεί οποιοσδήποτε χαρακτήρας του δοθέντος αλφάβητου, ή και θέσεις στις οποίες υπάρχει εναλλακτική επιλογή μεταξύ δύο χαρακτήρων. Παρ' όλα αυτά, δεν θα επεκταθούμε περισσότερο σε αυτό το θέμα, καθώς η παρούσα εργασία δεν περιλαμβάνει μελέτη μεθόδων που αφορούν τέτοιου είδους μοτίβα.

Την απλούστερη κατηγορία πιθανοτικού μοτίβου αποτελεί ο λεγόμενος *πίνακας βάρους θέσης* (position-weight matrix (PWM)) [2,8]. Αυτός αποτελεί μια στατιστική αναπαράσταση ενός μοτίβου (χωρίς κενά) μήκους K με χαρακτήρες από το αλφάβητο $A = \{a_1, a_2, ..., a_M\}$ υπό τη μορφή ενός πίνακα διάστασης MxK. Σε κάθε θέση αυτού του πίνακα, αντιστοιχεί η σχετική συχνότητα εμφάνισης του κάθε χαρακτήρα στη συγκεκριμένη θέση του μοτίβου. Για να γίνει πιο σαφής ο παραπάνω ορισμός (μια και θα χρησιμοποιηθεί στις περισσότερες μεθόδους που θα περιγράψουμε), θεωρούμε τις παρακάτω αλυσίδες:

1	5	8
↓	↓	↓
ABICEDDA		
CIDCA I EA		
AIBC I DAC		

με γράμματα από το αλφάβητο $A = \{A, B, C, D, E, I\}$ με $|A| = 6$. Τότε, ο αντίστοιχος *πίνακας θέσης βάρους* έχει προφανώς διάσταση $6x8$ και είναι ο ακόλουθος

θέση στο μοτίβο :

$$P = \begin{array}{c} \begin{array}{cccccccc} 1 & & & 4 & & & & 8 \\ \downarrow & & & \downarrow & & & & \downarrow \end{array} \\ \begin{array}{c} A \\ B \\ C \\ D \\ E \\ I \end{array} \begin{bmatrix} 2/3 & 0 & 0 & 0 & 1/3 & 0 & 1/3 & 2/3 \\ 0 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 1 & 0 & 0 & 0 & 1/3 \\ 0 & 0 & 1/3 & 0 & 0 & 2/3 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/3 & 0 & 1/3 & 0 \\ 0 & 2/3 & 1/3 & 0 & 1/3 & 1/3 & 0 & 0 \end{bmatrix} \\ \uparrow \\ \text{A (= αλφάβητο)} \end{array}.$$

Μάλιστα, παρατηρούμε ότι για κάθε στήλη $j \in \{1,2,...,8\}$ του πίνακα P, ισχύει $\sum_{i=1}^{6} p_{ij} = 1$ γεγονός που ήταν αναμενόμενο, αφού το άθροισμα των σχετικών συχνοτήτων εμφάνισης όλων των δυνατών γραμμάτων σε μία συγκεκριμένη θέση του μοτίβου ισούται με 1.

Ακόμα ένα μέτρο βαρύτητας μιας ακολουθίας θα μπορούσε να ορίζεται με μία δεδομένη *συνάρτηση βάρους* (που ασφαλώς παίρνει διαφορετικές τιμές για κάθε ακολουθία), καθώς και με το πλήθος των συμβολοσειρών στις οποίες εμφανίζεται ή με ένα συνδυασμό των δύο αυτών μέτρων.

Στην επόμενη παράγραφο, ακολουθεί ο ορισμός του ακριβούς προβλήματος, με το οποίο θα ασχοληθούμε.

1.1. Περιγραφή του Προβλήματος

Θεωρούμε, αρχικά, ένα πεπερασμένο σύνολο χαρακτήρων (συμβόλων) $A = \{a_1, a_2, ..., a_M\}$ (αλφάβητο) με $|A| = M$ και ένα σύνολο N συμβολοσειρών (ή αλυσίδων) $S = \{S_1, S_2, ..., S_N\}$ που έχει παραχθεί από το διακριτό αλφάβητο A. Στη γενική περίπτωση, υποθέτουμε μεταβλητού μήκους συμβολοσειρές, δηλαδή τέτοιες ώστε $|S_i| = L_i \geq K$, για

$i = 1,2,...,N$. Συμβολίζουμε, τώρα, με $m = [m_1 m_2 ... m_K]$ το προς αναζήτηση *μοτίβο* (πεπερασμένου μήκους K), όπου $m_j \in A$ για $j = 1,2,...,K$. Συγκεκριμένα, δεδομένου του μήκους K, στόχος μας είναι να ανακαλύψουμε μια κοινή ακολουθία χαρακτήρων (λέξη) $m = [m_1 m_2 ... m_K]$ μέσα στις αλυσίδες, η οποία εμφανίζεται μέσα σε αυτές με τη μεγαλύτερη δυνατή πιθανότητα. Πιο απλά, θέλουμε να βρούμε από όλες τις κοινές υπακολουθίες μήκους K μεταξύ των συμβολοσειρών, εκείνη που επαναλαμβάνεται (εμφανίζεται) περισσότερες φορές.

Για την πληρέστερη κατανόηση των παραπάνω, παρατίθεται το ακόλουθο σχήμα, όπου παρατηρούμε ότι ψάχνοντας για ένα μοτίβο μήκους 3, το επιθυμητό αποτέλεσμα θα ήταν η εύρεση της ακολουθίας χαρακτήρων "YOU". Αν, πάλι, αναζητούσαμε ένα μοτίβο μήκους 5, αυτό που εμφανίζεται με τη μεγαλύτερη συχνότητα μέσα στις συμβολοσειρές είναι η ακολουθία 'BETTER'.

```
                    1    5    10   15   20   25   30   35

S1 (L1 = 36) →  AIAYOUJUBETTERTHESTARSINTYOUDARKSKY
S2 (L2 = 28) →  YOUCANTAYITBETTERDENYILALALANGTHEM
S3 (L3 = 30) →  TRYBETTERAGEGETITIFYOUAREABLALALALTO
S4 (L4 = 31) →  IAMSUREYOUCANMAKEBETTERERSTAND
S5 (L5 = 36) →  LALALANEWANYOUTSTOMAKETOUFEELBETTER
```

Σχήμα 1.1: Συμβολοσειρές που περιέχουν τα μοτίβα $m_1 = [YOU]$ και $m_2 = [BETTER]$ μήκους 3 και 6 αντίστοιχα.

1.2. Η Μηχανική Μάθηση

Η Μηχανική Μάθηση αποτελεί έναν κλάδο της Πληροφορικής, ο οποίος μελετά αλγορίθμους που βασίζονται σε παρατηρούμενα δεδομένα και εφαρμόζονται σε ένα ευρύ φάσμα ερευνητικών περιοχών (Βιοπληροφορική, Αναγνώριση προτύπων κ.ά.). Κύριος στόχος της είναι να δημιουργήσει προγράμματα υπολογιστών που μέσα από την εμπειρική απόκτηση και την ενοποίηση γνώσεων (από μια συλλογή παρατηρήσεων που καλείται *σύνολο*

εκπαίδευσης) κατορθώνουν να βελτιώνονται συνεχώς. Ο βασικός λόγος, για τον οποίο κρίνεται απαραίτητη η περαιτέρω ανάπτυξή της, έγκειται στο ότι παρέχει τη δυνατότητα αυτόματης επίλυσης περίπλοκων προβλημάτων, τα οποία φαίνονται απρόσιτα στον ανθρώπινο νου, κυρίως λόγω του τεράστιου όγκου δεδομένων που πρέπει να χρησιμοποιηθούν για την επίλυσή τους. Ένα απλό παράδειγμα (που αφορά και το πρόβλημά μας) θα μπορούσε να είναι η ανακάλυψη ενός πολύπλοκου μοτίβου μέσα σε ένα πολύ μεγάλο σύνολο από αλυσίδες τεράστιου μήκους. Ένα τέτοιο πρόβλημα, ασφαλώς, μοιάζει απίθανο να λάβει άμεσης και έγκυρης απάντησης, ακόμα και από κάποιον εμπειρογνώμονα στο είδος, χωρίς τη χρήση κάποιου υπολογιστικού συστήματος. Σε τέτοιου είδους περιπτώσεις, λοιπόν, οδηγούμαστε υποχρεωτικά σε λύσεις που μας παρέχουν διάφορες μέθοδοι μηχανικής μάθησης, τις οποίες προσαρμόζουμε στο εκάστοτε πρόβλημα που μας απασχολεί.

Τα προβλήματα μηχανικής μάθησης κατατάσσονται σε τρεις κύριες κατηγορίες ανάλογα με την έξοδο που επιθυμούμε να δώσει ο αλγόριθμος που χρησιμοποιείται. Έτσι, διακρίνουμε τις κατηγορίες της *μάθησης με επίβλεψη* (supervised learning), της μάθησης *χωρίς επίβλεψη* (unsupervised learning) και της *ενισχυτικής* μάθησης (reinforcement learning). Στην πρώτη κατηγορία, ανήκουν οι περιπτώσεις αλγορίθμων που δημιουργούν μια συνάρτηση τέτοια, ώστε να αντιστοιχίζει μια συγκεκριμένη είσοδο σε κάποια αυστηρά καθορισμένη έξοδο. Κάτι τέτοιο επιτυγχάνεται μέσω ενός συνόλου εκπαίδευσης, αποτελούμενο από παραδείγματα ζευγών εισόδου και επιθυμητής εξόδου. Η μάθηση χωρίς επίβλεψη είναι μια μέθοδος που κυρίως ενδιαφέρεται να εκτιμήσει μια συνάρτηση κατανομής για το σύνολο εκπαίδευσης, βάσει συμπερασμάτων που εξάγει από αυτό, ενώ ασχολείται και με προβλήματα ομαδοποίησης. Εδώ, δεν γνωρίζουμε εξαρχής τη σωστή κατηγορία για κάποιο σύνολο παραδειγμάτων, αλλά προσπαθούμε να εκτιμήσουμε έναν κατάλληλο διαμερισμό των δεδομένων σε ομάδες. Γι' αυτό και τα προβλήματα αυτού του είδους είναι πιο δύσκολα. Στην περίπτωση, τέλος, της ενισχυτικής μάθησης, ο αλγόριθμος αποδίδει μια επιβράβευση (θετική ή και αρνητική), μετά από μια σειρά αποφάσεων. Σκοπός της είναι να μεγιστοποιήσει αυτή την επιβράβευση, η οποία εξαρτάται από το σύνολο αυτών των αποφάσεων, καθώς κάτι τέτοιο φανερώνει ότι αυτές μας οδήγησαν πολύ κοντά στον στόχο που μας ενδιαφέρει.

1.3. Δομή Διατριβής

Στα κεφάλαια που ακολουθούν, θα αναλύσουμε μια σειρά μεθόδων Μηχανικής Μάθησης που ανήκουν στη δεύτερη από τις προαναφερόμενες κατηγορίες μεθόδων (της παραγράφου 1.2) για την *προσδιορισμό ενός μοτίβου* μέσα σε ένα σύνολο συμβολοσειρών (ή αλυσίδων), οι οποίες αποτελούν το σύνολο των παρατηρήσεων, δίνοντας ως είσοδο το μήκος του και τις αλυσίδες (μέσα στις οποίες θα το αναζητήσουμε). Ο λόγος, για τον οποίο θα τις μελετήσουμε έγκειται στο γεγονός ότι στις περιπτώσεις που το μήκος του μοτίβου ή το πλήθος ή/και το μήκος των ακολουθιών είναι μεγάλο/α, το πρόβλημά μας γίνεται αρκετά περίπλοκο και η πρώτη κατηγορία μεθόδων δεν δύναται να μας δώσει γρήγορα μια λύση.

Συγκεκριμένα, στο Κεφάλαιο 2, θα περιγράψουμε τις προσεγγίσεις που βασίζονται στον αλγόριθμο EM (Expectation – Maximization) και τον Δειγματολήπτη Gibbs. Όπως θα διαπιστώσουμε προσεχώς, η ανίχνευση ενός μοτίβου μήκους K, σε αυτές τις μεθόδους πραγματοποιείται μέσω της ικανοποιητικής προσέγγισης ενός συνόλου παραμέτρων Ψ, μετά από ένα πλήθος επαναλήψεων. Επιπλέον, εξετάζουμε το αποτέλεσμα που παράγουν οι ίδιες μέθοδοι χρησιμοποιώντας κάποια είδη αρχικοποίησης, τα οποία παρέχονται κατόπιν χρήσης συγκεκριμένων αλγορίθμων ομαδοποίησης, όπως είναι οι Agglomerative (ιεραρχική ομαδοποίηση) και k-means (διαμεριστική ομαδοποίηση), τους οποίους θα παρουσιάσουμε στο Κεφάλαιο 3. Μια ακόμη σημαντική μέθοδο, την οποία θα μελετήσουμε, αποτελεί και ο αλγόριθμος Build-PST (ενισχυτική μάθηση) που κατασκευάζει ένα πιθανοτικό δέντρο προθεμάτων (περίπτωση *πιθανοτικού μοτίβου*), ο οποίος αναλύεται στο Κεφάλαιο 2. Το Κεφάλαιο 4 αποτελεί την καρδιά της παρούσας εργασίας, καθώς εισάγει δύο νέες μεθόδους, οι οποίες αποτελούν παραλλαγές του αλγορίθμου EM, προσπαθούν να παρέχουν μικρότερη εξάρτηση από την αρχικοποίηση και να προσφέρουν βέλτιστες λύσεις. Αυτές συγκρίνονται πειραματικά με τους μέχρι στιγμής υπάρχοντες αλγορίθμους (EM και Δειγματολήπτη Gibbs) στο Κεφάλαιο 5, με τυχαία και μη αρχικοποίηση και για λέξεις διαφορετικού μήκους, οι οποίες περιέχουν ή και όχι επαναλαμβανόμενα γράμματα ή συλλαβές. Τέλος, το Κεφάλαιο 6 περιλαμβάνει τον επίλογο της παρούσας εργασίας, καθώς επίσης τα θέματα προς συζήτηση και τα αναπάντητα ερωτήματα που ανακύπτουν από αυτήν.

ΚΕΦΑΛΑΙΟ 2. ΠΑΡΑΜΕΤΡΙΚΕΣ ΜΕΘΟΔΟΙ ΑΝΙΧΝΕΥΣΗΣ ΜΟΤΙΒΩΝ ΣΕ ΣΥΜΒΟΛΟΣΕΙΡΕΣ

Για την επίλυση του προς εξέταση προβλήματος, έχει επινοηθεί μια πληθώρα μεθόδων τα τελευταία χρόνια. Κάποιες από αυτές επιστρέφουν το βέλτιστο αποτέλεσμα με μεγάλο, όμως, κόστος όσον αφορά τον χρόνο που απαιτούν, όπως είναι η λεγόμενη «εξαντλητική μέθοδος» [8], ενώ κάποιες άλλες, σαφώς ταχύτερες, επιστρέφουν μια «προσεγγιστική λύση» με την έννοια ότι το επιστρεφόμενο αποτέλεσμα δεν είναι το καλύτερο δυνατό, αλλά ότι μέσα σε αυτό εμπεριέχεται ένα ικανοποιητικό ποσοστό χαρακτήρων από το ζητούμενο μοτίβο. Προφανώς, οι μέθοδοι που εμπίπτουν στην πρώτη κατηγορία είναι κατάλληλες μόνο για μικρά μοτίβα και χρονοτριβούν στην περίπτωση αναζήτησης μιας συμβολοσειράς μεγάλου μήκους ή όταν δίδεται ως είσοδος μεγάλος όγκος δεδομένων. Στην τελευταία κατηγορία, ανήκουν οι αλγόριθμοι ΕΜ (Expectation – Maximization) και Gibbs Sampling, τους οποίους θα παρουσιάσουμε αναλυτικά στο παρόν κεφάλαιο. Όπως θα διαπιστώσουμε και προσεχώς, η ανακάλυψη ενός μοτίβου μήκους K επιτυγχάνεται σε αυτές τις μεθόδους μέσω της ικανοποιητικής προσέγγισης ενός συνόλου παραμέτρων Ψ, μετά από ένα πλήθος επαναληπτικών εφαρμογών της εκάστοτε μεθόδου, και η κάθε μία από αυτές τερματίζει με βάση κάποιο κριτήριο σύγκλισης.

2.1. Το Πρόβλημα Εκτίμησης της Μέγιστης Πιθανοφάνειας

Έστω ένα σύνολο παρατηρήσεων $X = \{X_1, X_2, ..., X_r\}$ που τα στοιχεία του X_i είναι στατιστικά ανεξάρτητα μεταξύ τους και αποτελούν δείγματα μιας κατανομής με συνάρτηση πιθανότητας $P(X \mid \Theta)$, όπου Θ είναι ένα σύνολο παραμέτρων της κατανομής. Σε αυτή την περίπτωση, η συνάρτηση πιθανότητας για όλα τα στοιχεία X_i είναι προφανώς

$$P(X \mid \Theta) = \prod_{i=1}^{r} P(X_i \mid \Theta) = L(\Theta \mid X) \tag{2.1}$$

και καλείται *συνάρτηση Πιθανοφάνειας*. Από τον ορισμό της, γίνεται σαφές το γεγονός ότι πρόκειται για μία συνάρτηση, η οποία εξαρτάται άμεσα από τις παραμέτρους Θ, αφού εκφράζει την πιθανότητα των παρατηρήσεων X, γνωρίζοντας τις τιμές των παραμέτρων. Για να εκτιμήσουμε, λοιπόν, τη μέγιστη Πιθανοφάνεια, αρκεί να υπολογίσουμε για ποια τιμή των παραμέτρων Θ, η παραπάνω συνάρτηση λαμβάνει τη μέγιστη τιμή της

$$\Theta^* = \arg\max_{\Theta} L(\Theta \mid X). \tag{2.2}$$

2.2. Ο Αλγόριθμος ΕΜ για Μεικτές Κατανομές

Μια ευρέως γνωστή στατιστική μέθοδος που θεωρείται κατάλληλη για προβλήματα, στα οποία είτε το σύνολο δεδομένων είναι ελλιπές (δηλαδή απουσιάζουν κάποιες τιμές του) είτε υπάρχουν κρυμμένες μεταβλητές, είναι ο αλγόριθμος ΕΜ. Πρόκειται για μια επαναληπτική μέθοδο, η οποία αποσκοπεί στην «εκτίμηση της μέγιστης τιμής της συνάρτησης Πιθανοφάνειας» του εκάστοτε μοντέλου ούτως, ώστε να καθίσταται δυνατός ο υπολογισμός της τιμής των παραμέτρων του [4,14]. Εδώ, θα ασχοληθούμε με τη δεύτερη περίπτωση, καθώς σε αυτήν εντάσσεται το πρόβλημα που θα εξετάσουμε.

Αν θεωρήσουμε το σύνολο $Z = \{Z_1, Z_2, ..., Z_r\}$ που αποτελεί τις *κρυμμένες μεταβλητές* του μοντέλου, το πλήρες σύνολο δεδομένων είναι το $\{X, Z\}_{i=1}^{r}$. Έτσι, η λογαριθμική συνάρτηση πιθανοφάνειας δίνεται από τη σχέση (σύμφωνα με τον ορισμό της πιθανοφάνειας στην προηγούμενη παράγραφο)

$$\ln L(\Theta \mid X) = \ln P(X \mid \Theta) = \sum_{i=1}^{r} \ln P(X_i \mid \Theta), \tag{2.3}$$

Υποθέτουμε ότι τα δείγματα μπορούν να περιγραφούν από μια μεικτή κατανομή της μορφής

$$P(X_i \mid \Theta) = \sum_{j=1}^{w} \pi_j \cdot P(X_i \mid \theta_j), \tag{2.4}$$

όπου w το πλήθος των διαφορετικών κατανομών που υπάρχουν. Τότε, θεωρούμε το σύνολο των *παραμέτρων* $\Theta = \{\pi_1,...,\pi_w,\theta_1,...,\theta_w\}$, όπου $\{\theta_j\}_{j=1}^{w}$ είναι το σύνολο όλων των δυνατών κατανομών για τα στοιχεία του συνόλου X, π_j η εκ των προτέρων πιθανότητα ένα στοιχείο του X να ακολουθεί την κατανομή θ_j με $\sum_{j=1}^{w} \pi_j = 1$. Έτσι, κάθε κρυμμένη μεταβλητή Z_i λαμβάνει τιμές από το σύνολο $\{1,2,...,w\}$ και φανερώνει ποια από τις w διαφορετικές κατανομές ακολουθεί η αντίστοιχη παρατήρηση X_i.

Αντικαθιστώντας, λοιπόν, τη σχέση (2.4) στη (2.3) προκύπτει

$$\ln L(\Theta \mid X) = \sum_{i=1}^{r} \ln\left[\sum_{j=1}^{w} \pi_j \cdot P(X_i \mid \theta_j)\right]. \tag{2.5}$$

Επιπλέον, αν θεωρήσουμε την ποσότητα της λογαριθμικής πλήρους πιθανοφάνειας (δηλαδή εκείνης που προκύπτει από το πλήρες σύνολο δεδομένων), αυτή δίνεται από τον τύπο

$$\ln L(\Theta \mid X, Z) = \ln P(X, Z \mid \Theta), \tag{2.6}$$

όπου $P(X, Z \mid \Theta)$ είναι η από κοινού συνάρτηση πιθανότητας των μεταβλητών X και Z. Δεδομένου ότι οι παρατηρήσεις $\{X_i\}_{i=1}^{r}$ είναι ανεξάρτητες μεταξύ τους, έχουμε

$$P(X, Z \mid \Theta) = \prod_{i=1}^{r} P(X_i, Z_i \mid \Theta), \tag{2.7}$$

οπότε

$$\ln L(\Theta \mid X, Z) = \sum_{i=1}^{r} \ln P(X_i, Z_i \mid \Theta). \tag{2.8}$$

Επειδή, όμως, οι τιμές των μεταβλητών Z_i είναι άγνωστες και κάθε Z_i προσδιορίζει την κατανομή που ακολουθεί η παρατήρηση X_i, θα επιχειρήσουμε να υπολογίσουμε την παραπάνω ποσότητα χρησιμοποιώντας τον ορισμό της *δεσμευμένης πιθανότητας*, δηλαδή τη σχέση

$$P(X_i, Z_i \mid \Theta) = P(Z_i) \cdot P(X_i \mid Z_i, \Theta), \tag{2.9}$$

εκμεταλλευόμενοι το γεγονός ότι οι παρατηρήσεις $\{X_i\}_{i=1}^{r}$ είναι γνωστές.

Με βάση τη σχέση (2.8), η λογαριθμική πιθανοφάνεια παίρνει τη μορφή

$$\ln L(\Theta \mid X, Z) = \sum_{i=1}^{r} \ln\left[P(Z_i) \cdot P(X_i \mid Z_i, \Theta)\right] = \sum_{i=1}^{r} \ln\left[\pi_{Z_i} \cdot P(X_i \mid \theta_{Z_i})\right]. \tag{2.10}$$

Σε κάθε επανάληψη, ο αλγόριθμος ΕΜ υπολογίζει με βάση το Θεώρημα του Bayes, στο <u>*Expectation βήμα*</u>, τις *εκ των υστέρων* πιθανότητες

$$P(Z_i \mid X_i, \Theta) = \frac{P(Z_i) \cdot P(X_i \mid Z_i, \Theta)}{\sum_{l=1}^{w} P(Z_l) \cdot P(X_l \mid Z_l, \Theta)} = \frac{\pi_{Z_i} \cdot P(X_i \mid \theta_{Z_i})}{\sum_{l=1}^{w} \pi_{Z_l} \cdot P(X_l \mid \theta_{Z_l})}, i = 1,...,r. \tag{2.11}$$

Αφού, τώρα, η μέση (αναμενόμενη) τιμή της λογαριθμικής πιθανοφάνειας για τις παραμέτρους Θ (της τρέχουσας επανάληψης) είναι εξ' ορισμού

$$E\left[\ln L(\Theta \mid X, Z)\right] = E\left[\sum_{i=1}^{r} \ln\left(\pi_{Z_i} \cdot P(X_i \mid \theta_{Z_i})\right)\right] = \sum_{z \in Z} \sum_{i=1}^{r} \ln\left(\pi_{Z_i} \cdot P(X_i \mid \theta_{Z_i})\right) \cdot P(Z \mid X, \Theta)$$

$$= \sum_{z \in Z} \sum_{i=1}^{r} \ln\left(\pi_{Z_i} \cdot P(X_i \mid \theta_{Z_i})\right) \prod_{j=1}^{r} P(Z_j \mid X_j, \Theta)$$

$$= \sum_{Z_1=1}^{w} \sum_{Z_2=1}^{w} \ldots \sum_{Z_N=1}^{w} \sum_{i=1}^{r} \ln\left(\pi_{Z_i} \cdot P(X_i \mid \theta_{Z_i})\right) \prod_{j=1}^{r} P(Z_j \mid X_j, \Theta)$$

$$= \sum_{Z_1=1}^{w}\sum_{Z_2=1}^{w}\cdots\sum_{Z_N=1}^{w}\sum_{i=1}^{r}\sum_{j=1}^{w}\delta_{j,Z_i}\cdot\ln\left(\pi_j\cdot P(X_i\mid\theta_j)\right)\prod_{k=1}^{r}P(Z_k\mid X_k,\Theta)$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{w}\ln\left(\pi_j\cdot P(X_i\mid\theta_j)\right)\sum_{Z_1=1}^{w}\sum_{Z_2=1}^{w}\cdots\sum_{Z_N=1}^{w}\delta_{j,Z_i}\prod_{j=1}^{r}P(Z_k\mid X_k,\Theta)$$

και παρατηρώντας ότι, για $j=1,...,w,$ έχουμε

$$\sum_{Z_1=1}^{w}\sum_{Z_2=1}^{w}\cdots\sum_{Z_N=1}^{w}\delta_{j,Z_{ij}}\prod_{k=1}^{r}P(Z_k\mid X_k,\Theta)=\left(\sum_{Z_1=1}^{w}\cdots\sum_{Z_{i-1}=1}^{w}\sum_{Z_{i+1}=1}^{w}\cdots\sum_{Z_N=1}^{w}\prod_{k=1,k\neq i}^{r}P(Z_k\mid X_k,\Theta)\right)\cdot P(j\mid X_i,\Theta)$$

$$=\prod_{k=1,k\neq i}^{r}\left(\sum_{Z_j=1}^{w}P(Z_k\mid X_k,\Theta)\right)\cdot P(j\mid X_i,\Theta)=P(j\mid X_i,\Theta)\ ,$$

διότι ισχύει $\sum_{Z_j=1}^{w}P(Z_k\mid X_k,\Theta)=1$, οπότε προκύπτει ότι

$$E\left[\ln L(\Theta\mid X,Z)\right]=\sum_{i=1}^{r}\sum_{j=1}^{w}\ln\left(\pi_j\cdot P(X_i\mid\theta_j)\right)\cdot P(j\mid X_i,\Theta)$$

$$=\sum_{i=1}^{r}\sum_{j=1}^{w}\ln\left(\pi_j\right)\cdot P(j\mid X_i,\Theta)+\sum_{i=1}^{r}\sum_{j=1}^{w}\ln\left(P(X_i\mid\theta_j)\right)\cdot P(j\mid X_i,\Theta)\,. \qquad (2.12)$$

Στο δεύτερο βήμα του (το λεγόμενο *Maximization-βήμα*), ο αλγόριθμος ΕΜ υπολογίζει, με βάση τη μέση τιμή της λογαριθμικής πιθανοφάνειας (σχέση (2.12)) του προηγούμενου βήματος, τις νέες παραμέτρους $\Theta=\left\{\pi_j,\theta_j\right\}_{j=1}^{w}$, τις οποίες και χρησιμοποιεί στην επόμενη επανάληψη.

Πιο αναλυτικά, για να μεγιστοποιήσουμε την τελευταία συνάρτηση (διδόμενων των εκάστοτε παραμέτρων), θα μεγιστοποιήσουμε ξεχωριστά τους δύο όρους του αθροίσματος που περιέχει.

Εφόσον, λοιπόν, ισχύει ο περιορισμός $\sum_{j=1}^{w}\pi_j=1,$ χρησιμοποιούμε τον πολλαπλασιαστή Lagrange λ και θεωρούμε τη συνάρτηση

$$Q_{\pi} = \sum_{j=1}^{w}\sum_{i=1}^{r} \ln(\pi_j)\cdot P(j \mid X_i, \Theta) + \lambda \cdot \left(1 - \sum_{j=1}^{w} \pi_j\right). \tag{2.13}$$

Θέτοντας, τώρα, για κάποιο $j \in \{1,2,...,w\}$

$$\frac{\partial Q_{\pi}}{\partial \pi_j} = 0 \Rightarrow \frac{1}{\pi_j}\sum_{i=1}^{r} P(j \mid X_i, \Theta) = \lambda \Rightarrow \lambda \cdot \left(\sum_{j=1}^{w} \pi_j\right) = \sum_{i=1}^{r}\left(\sum_{j=1}^{w} P(j \mid X_i, \Theta)\right) \Rightarrow \lambda = r,$$

οπότε προκύπτουν (για την επόμενη επανάληψη του αλγορίθμου) οι νέες παράμετροι

$$\left\{\pi_j = \frac{1}{r}\left(\sum_{i=1}^{r} P(j \mid X_i, \Theta)\right)\right\}_{j=1}^{w}. \tag{2.14}$$

Ακόμα, λαμβάνοντας υπόψη τους περιορισμούς που ισχύουν για τις κατανομές $\{\theta_j\}_{j=1}^{w}$ και χρησιμοποιώντας, πάλι, πολλαπλασιαστές Lagrange, προκύπτουν οι νέες παράμετροι $\{\theta_j\}_{j=1}^{w}$ (προς χρήση στην επόμενη επανάληψη του αλγορίθμου).

Συμβολίζοντας, τώρα, τη μέση λογαριθμική πιθανοφάνεια στην $t-$στη επανάληψη του ΕΜ με $Q(\Theta, \Theta^{(t)})$, τα βήματα του αλγορίθμου συνοψίζονται ως εξής:

Πίνακας 2.1: Ο Αλγόριθμος ΕΜ.

- Αρχικά: Θεώρησε τυχαίες τιμές για τις παραμέτρους $\Theta^{(\pi)}$.
- Επανάλαβε τα παρακάτω
  - <u>Expectation – βήμα:</u> Υπολόγισε τις $P(Z_i \mid X_i, \Theta)$, $i = 1,...,r$
  - <u>Maximization – βήμα:</u> Εκτίμησε τις νέες παραμέτρους

$$\Theta^{(\nu)} = \arg\max_{\Theta} Q(\Theta, \Theta^{(\pi)}),$$

όπου

$$Q(\Theta, \Theta^{(\pi)}) = \sum_{z \in Z}\sum_{i=1}^{r} \ln\left(\pi_{Z_i} \cdot P(X_i \mid \theta_{Z_i})\right)\prod_{j=1}^{r} P(Z_j \mid X_j, \Theta^{(\pi)})$$

όσο ισχύει $Q(\Theta, \Theta^{(\nu)}) > Q(\Theta, \Theta^{(\pi)}) + \varepsilon$, όπου $\varepsilon > 0$.

Η εφαρμογή του αλγορίθμου ΕΜ, σε ένα δοθέν πρόβλημα που περιέχει ένα σύνολο παρατηρήσεων X και ένα σύνολο κρυμμένων μεταβλητών Z, εξασφαλίζει τη σύγκλιση σε κάποιο τοπικό μέγιστο της συνάρτησης πιθανοφάνειας $L(\Theta \mid X)$ μετά από ένα πλήθος επαναλήψεων των δύο βασικών βημάτων του, γεγονός που αποδεικνύεται στη συνέχεια.

Συμβολίζοντας ως $\ell(\Theta^{(t)})$ τη λογαριθμική συνάρτηση πιθανοφάνειας $\ln L(\Theta^{(t)} \mid X)$ στην $t-$στη επανάληψη του ΕΜ, σύμφωνα με το Θεώρημα του Bayes έχουμε

$$P(Z \mid X, \Theta^{(t+1)}) = \frac{P(X, Z \mid \Theta^{(t+1)})}{P(X \mid \Theta^{(t+1)})} \Rightarrow \ln P(Z \mid X, \Theta^{(t+1)}) = \ln P(X, Z \mid \Theta^{(t+1)}) - \ln P(X \mid \Theta^{(t+1)})$$

$$\Rightarrow \ln P(Z \mid X, \Theta^{(t+1)}) = \ln P(X, Z \mid \Theta^{(t+1)}) - \ell(\Theta^{(t+1)}) . \qquad (2.15)$$

Υπολογίζοντας, τώρα, τη μέση τιμή των δύο μελών της τελευταίας εξίσωσης ως προς την κατανομή $f(Z) = P(Z \mid X, \Theta^{(t)})$, λαμβάνουμε

$$E[\ln P(Z \mid X, \Theta^{(t+1)})] = E[\ln P(X, Z \mid \Theta^{(t+1)})] - \ell(\Theta^{(t+1)}) , \qquad (2.16)$$

διότι η $\ell(\Theta^{(t+1)}) = \ln L(\Theta^{(t+1)} \mid X)$ δεν εξαρτάται από τις μεταβλητές Z. Με τον ίδιο τρόπο, προκύπτει ότι στην $t-$στή επανάληψη του ΕΜ ισχύει

$$E[\ln P(Z \mid X, \Theta^{(t)})] = E[\ln P(X, Z \mid \Theta^{(t)})] - \ell(\Theta^{(t)}) . \qquad (2.17)$$

Αφαιρώντας κατά μέλη τις δύο τελευταίες εξισώσεις, προκύπτει τελικά

$$\ell(\Theta^{(t+1)}) - \ell(\Theta^{(t)}) = E[\ln P(X, Z \mid \Theta^{(t+1)})] - E[\ln P(X, Z \mid \Theta^{(t)})]$$

$$+ E[\ln P(Z \mid X, \Theta^{(t)})] - E[\ln P(Z \mid X, \Theta^{(t+1)})] . \qquad (2.18)$$

Λόγω, όμως, του ορισμού του Maximization βήματος του ΕΜ, ισχύει προφανώς

$$E[\ln P(X, Z \mid \Theta^{(t+1)})] - E[\ln P(X, Z \mid \Theta^{(t)})] = Q(\Theta, \Theta^{(t+1)}) - Q(\Theta, \Theta^{(t)}) \geq 0 , \qquad (2.19)$$

οπότε

$$\ell(\Theta^{(t+1)}) - \ell(\Theta^{(t)}) \geq E[\ln P(Z \mid X, \Theta^{(t)})] - E[\ln P(Z \mid X, \Theta^{(t+1)})]$$

$$= -\sum_{z \in Z} P(Z \mid X, \Theta^{(t)}) \cdot \ln\left(\frac{P(Z \mid X, \Theta^{(t+1)})}{P(Z \mid X, \Theta^{(t)})}\right) \geq 0, \qquad (2.20)$$

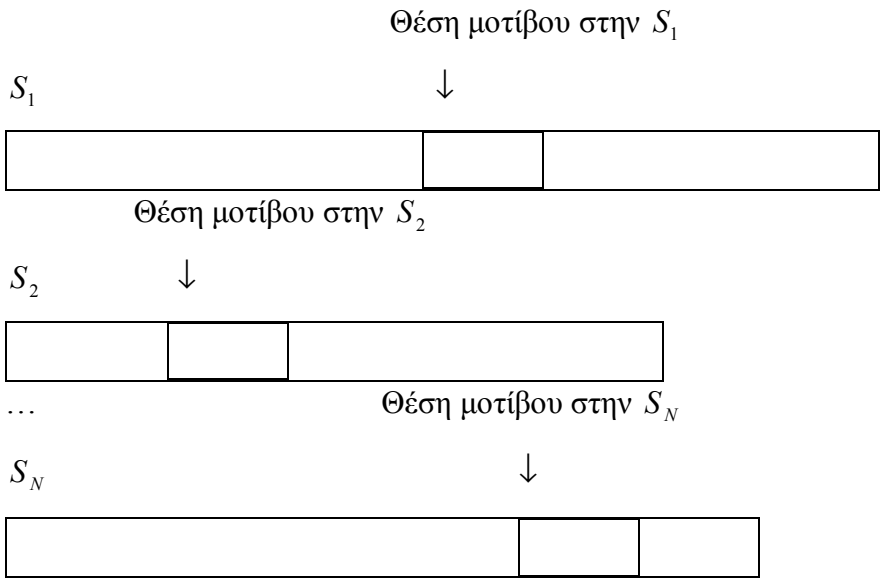
εφόσον εξ' ορισμού πρόκειται για την απόσταση Kullback-Leibler μεταξύ των συναρτήσεων πιθανότητας $P(Z \mid X, \Theta^{(t)})$ και $P(Z \mid X, \Theta^{(t+1)})$. Από τα παραπάνω έπεται ότι $\ell(\Theta^{(t+1)}) \geq \ell(\Theta^{(t)})$, δηλαδή η λογαριθμική συνάρτηση πιθανοφάνειας αυξάνεται σε κάθε βήμα του αλγορίθμου ΕΜ. Συνεπώς, από το σύνολο των επαναλήψεων του ΕΜ προκύπτει μια αύξουσα ακολουθία $\left(\ell(\Theta^{(t)})\right)_{t \geq 1}$, η οποία σύμφωνα με τον ορισμό της λογαριθμικής πιθανοφάνειας είναι άνω φραγμένη, δηλαδή παράγεται μια μονότονη και φραγμένη ακολουθία, γεγονός που αποδεικνύει ότι ο αλγόριθμος ΕΜ συγκλίνει σε κάποιο βήμα σε ένα τοπικό μέγιστο (local maximum) της λογαριθμικής συνάρτησης πιθανοφάνειας. Σε αυτό το σημείο, αξίζει να σημειώσουμε ότι αυτό είναι και το μειονέκτημα του ΕΜ, δηλαδή το γεγονός ότι η τυχαία αρχικοποίηση των παραμέτρων εξασφαλίζει τη σύγκλιση σε κάποιο τοπικό μέγιστο της συνάρτησης πιθανοφάνειας. Το ιδανικό θα ήταν, ασφαλώς, ο τερματισμός του αλγορίθμου στο ολικό μέγιστο (global maximum) της συνάρτησης πιθανοφάνειας, καθώς τότε μόνο θα λαμβάναμε την καλύτερη δυνατή λύση στο πρόβλημά μας.

Στην παράγραφο που ακολουθεί, θα κάνουμε μια σύντομη ανασκόπηση του εξεταζόμενου προβλήματος και θα εξηγήσουμε, αναλυτικά, με ποιον τρόπο εφαρμόζεται ο αλγόριθμος ΕΜ σε αυτό.

2.3. Εφαρμογή του ΕΜ στο Πρόβλημα της Ανίχνευσης Μοτίβου σε Συμβολοσειρές

Έστω ένα πεπερασμένο σύνολο χαρακτήρων (συμβόλων) $A = \{a_1, a_2, ..., a_M\}$ με $|A| = M$ και ένα σύνολο συμβολοσειρών $S = \{S_1, S_2, ..., S_N\}$, όπου κάθε στοιχείο του S_i αποτελείται από χαρακτήρες του αλφάβητου A και έχει μήκος $L_i \geq 1$, $i = 1,2,...,N$. Συμβολίζοντας με $m = [m_1 m_2 ... m_K]$ το προς αναζήτηση *μοτίβο* (πεπερασμένου μήκους K), όπου $m_j \in A$ για κάθε $j = 1,2,...,K$, παρατηρούμε ότι κάθε ακολουθία χαρακτήρων S_i περιέχει ακριβώς

$L_i - K + 1$ υπακολουθίες μήκους K (όσο το μήκος του μοτίβου), υπό την προϋπόθεση, βέβαια, ότι $L_i \geq K$. Συγκεντρώνοντας, τώρα, όλες τις υπακολουθίες μήκους K που προκύπτουν από τις συμβολοσειρές $S_1, S_2, ..., S_N$ στο σύνολο $X = \{X_1, X_2, ..., X_r\}$, διαπιστώνουμε ότι $r = \sum_{i=1}^{N}(L_i - K + 1) = \sum_{i=1}^{N} L_i - N(K-1)$. Μια γενική περιγραφή των παραπάνω, απεικονίζεται στο Σχήμα 2.1.



Σχήμα 2.1: Αλυσίδες χαρακτήρων που περιέχουν κάποιο μοτίβο καθορισμένου μήκους.

Το πρόβλημα που θα εξετάσουμε, όπως έχει ήδη αναφερθεί, αφορά την εύρεση των υπακολουθιών εκείνων, μέσα στο σύνολο X (*σύνολο εκπαίδευσης* του προβλήματος) που αποτελούν *μοτίβο*. Επιχειρώντας να το επιλύσουμε (παρόμοια περιγραφή συμπεριλαμβάνεται στις [1,2,4,5,7]), υποθέτουμε καταρχάς ότι κάθε υπακολουθία $X_i = X_{i_1} X_{i_2} ... X_{i_K}$ (με $X_{i_j} \in A$) ακολουθεί *πολυωνυμική κατανομή*, δηλαδή οι χαρακτήρες X_{i_j} (για $j \in \{1,2,...,K\}$) είναι ανεξάρτητοι μεταξύ τους σε κάθε θέση της X_i και ισχύει ότι

$$P(X_{i_j} = a_l^j) = p_{lj}, \quad \text{με} \quad \left|p_j\right| = \sum_{l=1}^{M} p_{lj} = 1. \tag{2.21}$$

Θεωρούμε, τώρα, ότι κάθε υπακολουθία X_i, $i = 1,2,...,r$, ή κάθε δυνατή λέξη του μοτίβου που παρήχθη από το σύνολο S μπορεί είτε να ανήκει στις υπακολουθίες εκείνες που αποτελούν μοτίβο, οπότε θα ισχύει ότι $X_i = m_1 m_2 ... m_K$ (εκτός κι αν υπάρχει μετάλλαξη σε

κάποιον/ους χαρακτήρα/ες) ή να μην αποτελεί μοτίβο και τότε, θα λέμε ότι ανήκει στο background.

Σύμφωνα με τα παραπάνω, το σύνολο X έχει παραχθεί από ένα *μεικτό μοντέλο πολυωνυμικών κατανομών* τέτοιο, ώστε κάθε υπακολουθία $X_i \in X$ να ανήκει είτε στο σύνολο των υπακολουθιών που αντιπροσωπεύουν το μοτίβο είτε στο σύνολο εκείνων που αποτελούν το background.

Στη συνέχεια, αν θεωρήσουμε μια «κρυμμένη» μεταβλητή Z_i για κάθε υπακολουθία X_i τέτοια, ώστε να ισχύει $Z_i = 1$, εάν η X_i ακολουθεί την κατανομή του μοτίβου, και $Z_i = 2$, αν η X_i ανήκει στο background, τότε η λύση του προβλήματος που θα εξετάσουμε παρέχεται άμεσα από τα ζεύγη $\{X_i, Z_i\}_{i=1}^{r}$ και συγκεκριμένα, από τις υπακολουθίες X_i, για τις οποίες ισχύει ότι Z_i=1. Εφόσον, όμως, δεν γνωρίζουμε την τιμή των μεταβλητών Z_i και το εν λόγω μοντέλο αποτελείται από δύο ειδών κατανομές (μοτίβο και background) , θεωρούμε ότι μια υπακολουθία ακολουθεί την κατανομή του μοτίβου με μια *εκ των προτέρων πιθανότητα* $\pi_1 = P(Z_i = 1)$ και την κατανομή του *background* με μια εκ των προτέρων πιθανότητα $\pi_2 = 1 - \pi_1 = P(Z_i = 2)$, διότι από τον ορισμό του μεικτού μοντέλου, πρέπει να ισχύει $\sum_{j=1}^{2} \pi_j = 1$. Επομένως, η συνάρτηση πιθανότητας για οποιαδήποτε υπακολουθία $X_i \in X$ δίδεται από τη σχέση

$$f(X_i \mid \Psi) = \pi_1 \cdot \rho(X_i \mid \theta) + (1 - \pi_1) \cdot \rho(X_i \mid b), \tag{2.22}$$

όπου $\Theta = \{\theta, b\}$ είναι ένα σύνολο των παραμέτρων για τις δύο πολυωνυμικές κατανομές, το οποίο συμπληρώνεται από το διάνυσμα $\pi = [\pi_1, \pi_2]$ και το συμβολίζουμε με $\Psi = \{\Theta, \pi\}$. Εδώ, αξίζει να τονίσουμε ότι το τελευταίο σύνολο παραμέτρων Ψ είναι αυτό, το οποίο θα επιχειρήσει ο αλγόριθμος ΕΜ να εκτιμήσει. Συγκεκριμένα, $\theta = [\theta_{jk}]$ είναι ένας πίνακας διάστασης MxK τέτοιος, ώστε κάθε στοιχείο του θ_{jk} αντιπροσωπεύει την πιθανότητα να εμφανίζεται ο χαρακτήρας a_j στην $k-$στη θέση του μοτίβου και ο οποίος πληροί τον περιορισμό $\sum_{j=1}^{M} \theta_{jk} = 1$, για κάθε $k \in \{1,2,...,K\}$, και $b = [b_i]$ ένα διάνυσμα διάστασης M με

$\sum_{i=1}^{M} b_i = 1$, όπου το στοιχείο b_i παριστάνει την πιθανότητα εμφάνισης του χαρακτήρα $a_i \in A\ (i = 1,2,...,M)$ στις υπακολουθίες που απαρτίζουν το *background*. Αφού, επιπλέον, έχουμε υποθέσει ότι οι θέσεις σε κάθε υπακολουθία είναι ανεξάρτητες μεταξύ τους, οι συναρτήσεις κατανομής για το μοντέλο του *μοτίβου* και του *background* είναι αντίστοιχα

$$\rho(X_i \mid \theta) = P(X_i \mid Z_i = 1, \Theta) = \prod_{k=1}^{K} \theta_{j,X_{i_k}} = \prod_{k=1}^{K} \prod_{j=1}^{M} \theta_{jk}^{\delta(X_{i_k}, a_j)} \tag{2.23}$$

και

$$\rho(X_i \mid b) = P(X_i \mid Z_i = 0, \Theta) = \prod_{k=1}^{K} b_j = \prod_{k=1}^{K} \prod_{j=1}^{M} b_j^{\delta(X_{i_k}, a_j)} = \prod_{j=1}^{M} b_j^{\sum_{k=1}^{K} \delta(X_{i_k}, a_j)}, \tag{2.24}$$

όπου με δ συμβολίζουμε την δείκτρια συνάρτηση $\delta(X_{i_k}, a_j) = \begin{cases} 1, X_{i_k} = a_j \\ 0, X_{i_k} \neq a_j \end{cases}$.

Σε αυτή την περίπτωση, αυτό που επιθυμούμε είναι να μεγιστοποιήσουμε τη λογαριθμική συνάρτηση (πιθανοφάνειας)

$$L(X) = \ln P(X \mid \Psi) = \sum_{i=1}^{r} \ln f(X_i \mid \Psi)$$

$$= \sum_{i=1}^{r} \ln\left[\pi_1 \cdot \rho(X_i \mid \theta) + (1 - \pi_1) \cdot \rho(X_i \mid b)\right]. \tag{2.25}$$

Θέτοντας, τώρα, $\Phi_1(X_i \mid \Theta) = \rho(X_i \mid \theta)$ και $\Phi_2(X_i \mid \Theta) = \rho(X_i \mid b)$, στο Expectation – βήμα του, ο αλγόριθμος ΕΜ υπολογίζει τις εκ των υστέρων πιθανότητες

$$P(Z_i = j \mid X_i) = \frac{\pi_j \cdot \Phi_j(X_i \mid \Theta)}{\sum_{l=1}^{2} \pi_l \cdot \Phi_l(X_i \mid \Theta)}, j = 1,2, \tag{2.26}$$

διότι εδώ έχουμε δύο μόνο διαφορετικές κατανομές που μπορούν να ακολουθούν οι παρατηρήσεις μας. Η μέση (πλήρης) λογαριθμική πιθανοφάνεια, την οποία θα μεγιστοποιήσουμε, είναι τώρα

$$Q = \sum_{i=1}^{r} \left[\ln P(X_i, Z_i \mid \Psi)\right] \cdot P(Z_i \mid X_i, \Psi) = \sum_{i=1}^{r}\sum_{j=1}^{2} P(Z_i = j \mid X_i, \Psi) \cdot \ln P(X_i, Z_i = j \mid \Psi)$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{2} P(Z_i = j \mid X_i, \Psi) \cdot \ln\left[P(Z_i = j) \cdot P(X_i \mid Z_i = j, \Psi)\right]$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{2} P(Z_i = j \mid X_i, \Psi) \cdot \ln\left[\pi_j \cdot \Phi_j(X_i \mid \Theta)\right] \Rightarrow$$

$$Q = \sum_{i=1}^{r}\sum_{j=1}^{2} P(Z_i = j \mid X_i, \Psi) \cdot \ln \pi_j + \sum_{i=1}^{r}\sum_{j=1}^{2} P(Z_i = j \mid X_i, \Psi) \cdot \ln \Phi_j(X_i \mid \Theta)\,. \qquad (2.27)$$

Συνεχίζοντας στο <u>*Maximization-βήμα*</u>, ο αλγόριθμος ΕΜ υπολογίζει βασιζόμενος στην παραπάνω σχέση για την Q, τις νέες παραμέτρους $\Psi = \left\{(\pi_j)_{j=1}^{2}, \theta, b\right\}$, τις οποίες και χρησιμοποιεί στην επόμενη επανάληψη. Όπως ακριβώς και στην προηγούμενη παράγραφο (εδώ, όμως, για $w = 2$), οι νέες εκ των προτέρων πιθανότητες δίνονται από τη σχέση

$$\pi_j = \frac{1}{r}\left(\sum_{i=1}^{r} P(Z_i = j \mid X_i, \Psi)\right),\ j = 1,2\,. \qquad (2.28)$$

Αφού, τώρα, μας είναι γνωστές οι συναρτήσεις πιθανότητας για τις δύο κατανομές που ακολουθούν οι παρατηρήσεις, θα περιγράψουμε πιο αναλυτικά πώς προκύπτουν οι σχέσεις για τις νέες παραμέτρους $\Theta = \{\theta, b\}$. Χρησιμοποιώντας τον δεύτερο όρο του αθροίσματος της σχέσης (2.27), τον θέτουμε ως $Q_\Phi = Q_\theta + Q_b$, όπου

$$Q_\theta = \sum_{i=1}^{r} P(Z_i = 1 \mid X_i, \Psi) \cdot \ln\left[\prod_{k=1}^{K}\prod_{j=1}^{M} \theta_{jk}^{\,\delta(X_{i_k}, a_j)}\right]$$

$$= \sum_{i=1}^{r} P(Z_i = 1 \mid X_i, \Psi) \cdot \sum_{k=1}^{K}\sum_{j=1}^{M} \delta(X_{i_k}, a_j) \cdot \ln(\theta_{jk}) \qquad (2.29)$$

και

$$Q_b = \sum_{i=1}^{r} P(Z_i = 2 \mid X_i, \Psi) \cdot \ln\left[\prod_{k=1}^{K} \prod_{j=1}^{M} b_j^{\,\delta(X_{i_k}, a_j)} \right]$$

$$= \sum_{i=1}^{r} P(Z_i = 2 \mid X_i, \Psi) \cdot \sum_{k=1}^{K} \sum_{j=1}^{M} \delta(X_{i_k}, a_j) \cdot \ln(b_j). \qquad (2.30)$$

Προκειμένου, λοιπόν, να υπολογίσουμε τον καινούριο πίνακα $\theta = [\theta_{jk}]$, λαμβάνουμε υπόψη τους περιορισμούς $\sum_{j=1}^{M} \theta_{jk} = 1$, για $k = 1,2,...,K$. Εν συνεχεία, θεωρούμε τους πολλαπλασιαστές Lagrange λ_k, $k = 1,2,...,K$, και παραγωγίζοντας ως προς θ_{jk} (και θέτοντας ίση με μηδέν) τη συνάρτηση

$$Q_1 = \sum_{i=1}^{r} P(Z_i = 1 \mid X_i, \Psi) \cdot \sum_{k=1}^{K} \sum_{j=1}^{M} \delta(X_{i_k}, a_j) \cdot \ln(\theta_{jk}) + \sum_{k=1}^{K} \lambda_k \cdot \left(1 - \theta_{jk}\right), \qquad (2.31)$$

προκύπτει ότι

$$\sum_{i=1}^{r} P(Z_i = 1 \mid X_i, \Psi) \cdot \delta(X_{i_k}, a_j) \cdot \frac{1}{\theta_{jk}} - \lambda_k = 0$$

$$\Rightarrow \lambda_k \cdot \sum_{j=1}^{M} \theta_{jk} = \sum_{j=1}^{M} \sum_{i=1}^{r} P(Z_i = 1 \mid X_i, \Psi) \cdot \delta(X_{i_k}, a_j)$$

$$\Rightarrow \lambda_k = \sum_{i=1}^{r} P(Z_i = 1 \mid X_i, \Psi) \sum_{j=1}^{M} \delta(X_{i_k}, a_j)$$

$$\Rightarrow \lambda_k = \sum_{i=1}^{r} P(Z_i = 1 \mid X_i, \Psi), \qquad (2.32)$$

διότι ισχύει $\sum_{j=1}^{M} \delta(X_{i_k}, a_j) = 1$, για κάθε $i = 1,...,r$, $k = 1,...,K$.

Με βάση τα παραπάνω, ο νέος πίνακας $\theta = [\theta_{jk}]$ δίνεται από τις σχέσεις

$$\theta_{jk} = \frac{\sum_{i=1}^{r} P(Z_i = 1 \mid X_i, \Psi) \cdot \delta(X_{i_k}, a_j)}{\sum_{i=1}^{r} P(Z_i = 1 \mid X_i, \Psi)}, j = 1,...,M,\ k = 1,...,K\ . \tag{2.33}$$

Για τον υπολογισμό, τώρα, του νέου διανύσματος b, θεωρούμε τον πολλαπλασιαστή Lagrange λ (λόγω του περιορισμού $\sum_{j=1}^{M} b_j = 1$) και παραγωγίζοντας ως προς b_j (και θέτοντας ίση με μηδέν) τη συνάρτηση

$$Q_2 = \sum_{i=1}^{r} P(Z_i = 2 \mid X_i, \Psi) \cdot \sum_{k=1}^{K} \sum_{j=1}^{M} \delta(X_{i_k}, a_j) \cdot \ln(b_j) - \lambda \cdot \left(1 - \sum_{j=1}^{M} b_j \right), \tag{2.34}$$

παίρνουμε

$$\lambda = \sum_{i=1}^{r} P(Z_i = 2 \mid X_i, \Psi) \cdot \sum_{k=1}^{K} \delta(X_{i_k}, a_j) \cdot \frac{1}{b_j}$$

$$\Rightarrow \lambda \cdot \sum_{j=1}^{M} b_j = \sum_{i=1}^{r} P(Z_i = 2 \mid X_i, \Psi) \cdot \sum_{j=1}^{M} \sum_{k=1}^{K} \delta(X_{i_k}, a_j)$$

$$\Rightarrow \lambda = K \cdot \sum_{i=1}^{r} P(Z_i = 2 \mid X_i, \Psi), \tag{2.35}$$

διότι εξ' ορισμού είναι $\sum_{j=1}^{M} \sum_{k=1}^{K} \delta(X_{i_k}, a_j) = K$.

Επομένως, το καινούριο διάνυσμα b δίνεται από τις σχέσεις

$$b_j = \frac{\sum_{i=1}^{r} P(Z_i = 2 \mid X_i, \Psi) \cdot \sum_{k=1}^{K} \delta(X_{i_k}, a_j)}{K \cdot \sum_{i=1}^{r} P(Z_i = 2 \mid X_i, \Psi)}, j = 1,...,M. \tag{2.36}$$

Συνοψίζοντας τα παραπάνω αποτελέσματα, ο αλγόριθμος ΕΜ παίρνει την εξής μορφή για το πρόβλημα που εξετάζουμε:

Πίνακας 2.2: Ο Αλγόριθμος ΕΜ για το Πρόβλημα.

- Αρχικά: Θεώρησε τυχαίες τιμές για τις παραμέτρους $\Psi^{(\pi)}$.
- Επανάλαβε τα παρακάτω
  - <u>*Expectation – βήμα:*</u> Υπολόγισε τις εκ των υστέρων πιθανότητες

$$P(Z_i = j \mid X_i) = \frac{\pi_j \cdot \Phi_j(X_i \mid \Theta)}{\sum_{l=1}^{2} \pi_l \cdot \Phi_l(X_i \mid \Theta)},\ i = 1,...,r,\ j = 1,2.$$

  - <u>*Maximization – βήμα:*</u> Εκτίμησε τις νέες παραμέτρους $\Psi^{(\nu)} = \arg\max_{\Psi} Q(\Psi, \Psi^{(\pi)})$,

όπου

$$Q(\Psi, \Psi^{(\pi)}) = \sum_{i=1}^{r} \sum_{j=1}^{2} P(Z_i = j \mid X_i, \Psi) \cdot \ln P(X_i, Z_i = j \mid \Psi),$$

από τις σχέσεις:

$$1)\quad \pi_j = \frac{1}{r}\left(\sum_{i=1}^{r} P(Z_i = j \mid X_i, \Psi) \right),\ j = 1,2,$$

$$2)\quad \theta_{jk} = \frac{\sum_{i=1}^{r} P(Z_i = 1 \mid X_i, \Psi) \cdot \delta(X_{i_k}, a_j)}{\sum_{i=1}^{r} P(Z_i = 1 \mid X_i, \Psi)},\ j = 1,...,M,\ k = 1,...,K,$$

$$3)\quad b_j = \frac{\sum_{i=1}^{r} P(Z_i = 2 \mid X_i, \Psi) \cdot \sum_{k=1}^{K} \delta(X_{i_k}, a_j)}{K \cdot \sum_{i=1}^{r} P(Z_i = 2 \mid X_i, \Psi)},\ j = 1,...,M,$$

όσο ισχύει $Q(\Psi, \Psi^{(\nu)}) > Q(\Psi, \Psi^{(\pi)}) + \varepsilon$, όπου $\varepsilon > 0$.

<u>Παρατήρηση:</u> Αφού εφαρμόσουμε τον αλγόριθμο ΕΜ, λαμβάνουμε ως έξοδο τον πίνακα βάρους θέσης $\theta = [\theta_{jk}] \in \Re^{MxK}$ του μοτίβου, καθώς και ένα διάνυσμα πιθανοτήτων $b = [b_1, b_2, ..., b_M] \in \Re^M$, της κατανομής του background. Ο παραπάνω πίνακας θ αποτελεί τη λύση που δίνει ο αλγόριθμος ΕΜ για το πρόβλημα που εξετάζουμε, αφού λαμβάνοντας το μέγιστο στοιχείο κάθε στήλης του και βρίσκοντας σε ποιο γράμμα αυτό αντιστοιχεί, θα προκύψει το πιο πιθανό μοτίβο μήκους K που θέλουμε να προσδιορίσουμε. Τέλος, οι υπακολουθίες που σύμφωνα με τον αλγόριθμο αναπαριστούν το μοτίβο, είναι αυτές που

μεγιστοποιούν τις εκ των υστέρων πιθανότητες του Expectation – βήματος για την κατανομή του μοτίβου.

Παράδειγμα: Έστω ότι διαθέτουμε το σύνολο χαρακτήρων $\Omega = \{A, B, I, E, G\}$ και αναζητούμε το μοτίβο $m = [BIG]$ του προηγούμενου παραδείγματος μέσα σε ένα σύνολο X από συμβολοσειρές μήκους 3 με $|X| = 20$, στις οποίες υπάρχουν 5 αντίγραφα του μοτίβου, με πιθανότητα μετάλλαξης κάποιου χαρακτήρα 20%, και αυτά εμπεριέχονται στο σύνολο $D = \{BIG, IIG, BIE, AIG, BIB\}$.

Τότε, ο πίνακας θέσης βάρους που προκύπτει (από το σύνολο D) είναι

$$\theta = \begin{array}{c} \\ A \\ B \\ I \\ E \\ G \end{array} \begin{bmatrix} 1/5 & 0 & 0 \\ 3/5 & 0 & 1/5 \\ 1/5 & 1 & 0 \\ 0 & 0 & 1/5 \\ 0 & 0 & 3/5 \end{bmatrix},$$

όπου παρατηρούμε ότι $\sum_{j=1}^{5} \theta_{jk} = 1$, για κάθε $k \in \{1,2,3\}$.

Αν, ακόμη, υποθέσουμε ότι είναι $X = \{D, ABI, IIB, ABB, AIE, GAE, EGG, EEA, AEI, IGA, GBA, BAE, BEE, GIB, AGB, IAG\}$, η παράμετρος που αντιστοιχεί στην κατανομή του background είναι το διάνυσμα

$$b = [12/60 \quad 13/60 \quad 14/60 \quad 10/60 \quad 11/60],$$

για το οποίο πράγματι ισχύει

$$\sum_{i=1}^{5} b_i = \frac{12+13+14+10+11}{60} = \frac{60}{60} = 1 .$$

Ο παραπάνω πίνακας θ και το διάνυσμα b αποτελούν την καλύτερη δυνατή λύση που θα μπορούσε να προκύψει από τον αλγόριθμο ΕΜ για το σύνολο παρατηρήσεων X. Πράγματι, βρίσκοντας το μέγιστο στοιχείο κάθε στήλης του πίνακα θ και παρατηρώντας κατά σειρά ότι, για την πρώτη στήλη αυτό αντιστοιχεί στο γράμμα 'B', για την δεύτερη στήλη στον χαρακτήρα 'I' και για την τρίτη στήλη στο γράμμα 'G', παρατηρούμε ότι το *μοτίβο* που προκύπτει από τον πίνακα θ είναι η ακολουθία χαρακτήρων "BIG".

2.4. Ο Αλγόριθμος Gibbs Sampling

Μια γενική στοχαστική μέθοδος, η οποία αποβλέπει στον καθορισμό των παραμέτρων ενός στατιστικού μοντέλου που σχετίζεται με ένα σύνολο δεδομένων, είναι ο αλγόριθμος Gibbs Sampling. Βασίζεται στη στατιστική μέθοδο της επαναληπτικής δειγματοληψίας και αποτελεί μια απλουστευμένη μορφή της μεθόδου MCMC (Markov chain Monte Carlo) [9,18]. Η τελευταία μέθοδος χρησιμοποιείται στην περίπτωση που έχουμε μια πολυδιάστατη τυχαία μεταβλητή $Y = (Y_1,...,Y_n)$ με κατανομή $f(Y)$, από την οποία θέλουμε να πάρουμε κάποια δείγματα, γεγονός όμως που δεν είναι εφικτό λόγω του ότι η f είναι εξαιρετικά πολύπλοκη. Η λύση παρέχεται, τότε, μέσω της MCMC, σύμφωνα με την οποία λαμβάνουμε δείγματα για κάθε μονοδιάστατη τυχαία μεταβλητή Y_i από την περιθώρια συνάρτηση κατανομής της και η διαδικασία αυτή επαναλαμβάνεται, έως ότου οδηγηθούμε σε μια στάσιμη κατανομή για την Y.

Ανάλογα, τώρα, ενεργεί και ο Gibbs Sampling [21]. Έστω ότι έχουμε ένα σύνολο παρατηρήσεων X και ένα σύνολο παραμέτρων Ψ. Ακόμα, έστω θ το σύνολο των παραμέτρων που θέλουμε να εκτιμήσουμε. Ξεκινώντας με μια τυχαία τιμή για τις Ψ, σε κάθε βήμα λαμβάνουμε ένα δείγμα για κάθε μία από τις παραμέτρους που μας ενδιαφέρουν, διδόμενων των τιμών των υπολοίπων παραμέτρων και των παρατηρήσεων. Η διαδικασία

αυτή επαναλαμβάνεται μέχρι τη στιγμή που οι παράμετροι $\theta = (\theta_1, \theta_2, ..., \theta_D)$ δεν αλλάζουν, οπότε ο αλγόριθμος τερματίζει.

Στη συνέχεια, ακολουθεί η γενική περιγραφή του αλγορίθμου υπό τη μορφή ψευδοκώδικα.

Πίνακας 2.3: Ο Αλγόριθμος Gibbs Sampling.

- Διάλεξε τυχαία τις αρχικές τιμές των παραμέτρων $\Psi^{(0)}$ και $\theta^{(0)}$.
- Επανάλαβε τα παρακάτω (στην $i-$στή επανάληψη του αλγορίθμου) :
  - Πάρε ένα δείγμα για την $\theta_1^{(i)}$ από την $\rho(\theta_1 \mid \theta_2^{(i-1)}, ..., \theta_D^{(i-1)}, \Psi^{(i-1)}, X)$.
  - Πάρε ένα δείγμα για την $\theta_2^{(i)}$ από την $\rho(\theta_2 \mid \theta_1^{(i-1)}, \theta_3^{(i-1)} ..., \theta_D^{(i-1)}, \Psi^{(i-1)}, X)$.
  - …
  - Πάρε ένα δείγμα για την $\theta_D^{(i)}$ από την $\rho(\theta_D \mid \theta_1^{(i-1)}, ..., \theta_{D-1}^{(i-1)}, \Psi^{(i-1)}, X)$.
  - Πάρε ένα δείγμα για το $\Psi^{(i)}$ από την $\rho(\Psi \mid \theta_1^{(i-1)}, ..., \theta_D^{(i-1)}, X)$

έως ότου καταλήξεις σε μια στάσιμη κατάσταση.

Στον Πίνακα 2.3, οι $\theta^{(0)}, \theta^{(1)}, ..., \theta^{(t)}$ αναπαριστούν την πραγματοποίηση μιας αλυσίδας Markov και συγκεκριμένα, η πιθανότητα μετάβασης από την κατάσταση $\theta^{(i-1)}$ στη $\theta^{(i)}$ υπολογίζεται από τη σχέση

$$T(\theta^{(i-1)}, \theta^{(i)}) = \rho(\theta_1^{(i)} \mid \theta_2^{(i-1)}, ..., \theta_D^{(i-1)}, \Psi^{(i-1)}, X) \cdot \rho(\theta_2^{(i)} \mid \theta_1^{(i)}, \theta_3^{(i-1)}, ..., \theta_D^{(i-1)}, \Psi^{(i-1)}, X) \cdot ... \cdot$$

$$\rho(\theta_D^{(i)} \mid \theta_1^{(i)}, , ..., \theta_D^{(i)}, \Psi^{(i-1)}, X).$$

Με τη διαδικασία που περιγράφει ο αλγόριθμος, υπονοεί ότι η από κοινού κατανομή των $(\theta_1^{(t)},...,\theta_D^{(t)},\Psi^{(t)})$ μετά από κάποιο πλήθος επαναλήψεων θα συγκλίνει στην κατανομή $\rho(\theta_1,...,\theta_D,\Psi \mid X)$.

2.5. Εφαρμογή του Αλγορίθμου Gibbs Sampling στο Εξεταζόμενο Πρόβλημα

Θεωρούμε, και πάλι, το γνωστό μας πρόβλημα με σύνολο δεδομένων το σύνολο των συμβολοσειρών $S=\{S_1,S_2,...,S_N\}$ με γράμματα από το αλφάβητο $A=\{a_1,a_2,...,a_M\}$, μέσα στις οποίες αναζητούμε ένα μοτίβο καθορισμένου μήκους K (που αποτελεί το στατιστικό μοντέλο του προβλήματος) και όπου κάθε αλυσίδα S_i έχει συγκεκριμένο (γνωστό) μήκος $L_i \geq K$, για κάθε $i \in \{1,2,...,N\}$. Έτσι, σε κάθε συμβολοσειρά S_i υπάρχουν L_i-K+1 δυνατές θέσεις έναρξης του μοτίβου. Όπως παρατηρούμε, το πρόβλημα που εξετάζουμε θα μπορούσε να νοηθεί ως αναζήτηση μιας τοπικής καθορισμένου μήκους ευθυγράμμισης των αλυσίδων. Στόχος της μεθόδου που εξετάζουμε, στη συγκεκριμένη περίπτωση, είναι η εύρεση του πιο πιθανού κοινού μοτίβου μεταξύ των δοσμένων συμβολοσειρών.

Στην περίπτωση εφαρμογής του αλγορίθμου Gibbs Sampling [15,22,24], το σύνολο των παρατηρήσεων του προβλήματος αποτελείται από το σύνολο $X=\{X_1,X_2,...,X_r\}$ των υπακολουθιών μήκους K που προκύπτουν από κάθε συμβολοσειρά ή ισοδύναμα από όλες τις δυνατές θέσεις έναρξης $\{d_i\}_{i=1}^{r}$ του μοτίβου για κάθε μία από τις δοθείσες αλυσίδες, ενώ ως κρυμμένες μεταβλητές θεωρούμε τη θέση έναρξης του μοτίβου σε κάθε αλυσίδα. Οι παράμετροι Ψ, τώρα, που αναζητούμε και οι οποίες αλλάζουν μετά από κάθε επανάληψη του αλγορίθμου, είναι ο πίνακας βάρους θέσης θ, διάστασης MxK, που αντιστοιχεί στο ευρισκόμενο κάθε φορά μοτίβο και το διάνυσμα b, διάστασης M, το οποίο αντιστοιχεί στο background του προβλήματος. Ως κριτήριο τερματισμού του αλγορίθμου, θεωρούμε την περίπτωση που σε κάποια επανάληψη οι θέσεις μοτίβου δεν αλλάξουν ή διαφορετικά την περίπτωση, στην οποία ο πίνακας βάρους θέσης του μοτίβου συγκλίνει. Οι βοηθητικές παράμετροι $\Theta=\{\Theta_1,\Theta_2,...,\Theta_r\}$ αποτελούνται από το σύνολο των ποσοτήτων $\{P(X_i \mid \theta),P(X_i \mid b)\}_{i=1}^{r}$, όπου με $P(X_i \mid \theta)$ συμβολίζουμε την πιθανότητα η υπακολουθία X_i να είναι μοτίβο, ενώ με $P(X_i \mid b)$ την πιθανότητα η X_i να ανήκει στο background. Οι

πιθανότητες $\{P(X_i \mid \theta), P(X_i \mid b)\}_{i=1}^{r}$ υπολογίζονται όπως και στην περίπτωση του αλγορίθμου ΕΜ. Η μόνη διαφορά που υπάρχει αφορά τον υπολογισμό των παραμέτρων στα επαναληπτικά βήματα. Συγκεκριμένα, συμβολίζοντας με c_{0j} το πλήθος εμφανίσεων του χαρακτήρα a_i $(i = 1,2,...,M)$ στο background, c_{ij} το πλήθος εμφανίσεων του χαρακτήρα a_i $(i = 1,2,...,M)$ στη θέση j του μοτίβου $(j = 1,...,K)$ και με r_j $(j = 1,2,...,M)$ κάποιες ψευδομετρήσεις που αντιστοιχούν σε κάθε γράμμα του αλφάβητου με $R = \sum_{j=1}^{M} r_j$ το σύνολο αυτών των ψευδομετρήσεων (που μπορούμε να θεωρήσουμε, για παράδειγμα, ότι αντιστοιχούν στο 10% των μετρήσεων που αντιστοιχεί σε κάθε γράμμα για όλες τις θέσεις των αλυσίδων), τότε οι παράμετροι θ και b του προβλήματος δίνονται σε κάθε βήμα του αλγορίθμου, αντίστοιχα, από τους τύπους:

$$\theta_{ij} = \frac{c_{ij} + r_j}{N - 1 + R} \tag{2.37}$$

και

$$b_i = \frac{c_{0i} + r_i}{\sum_{k=1}^{M} c_{0k} + R}, \tag{2.38}$$

όπου $i \in \{1,2,...,M\}$, $j \in \{1,2,...,k\}$. Μάλιστα, ο λόγος για τον οποίο χρησιμοποιούμε τις ψευδομετρήσεις έγκειται στην αποφυγή ισότητας με το μηδέν για κάποιο στοιχείο του MxK πίνακα θ ή του $M-$διάστατου διανύσματος b. Θέτοντας, τώρα, $Q_i = \frac{P(X_i \mid \theta)}{P(X_i \mid b)}$ για την υπακολουθία X_i, όπως θα διαπιστώσουμε παρακάτω οι ποσότητες $\{Q_i\}_{i=1}^{r}$ αποτελούν το βασικό μέτρο βαρύτητας για κάθε υπακολουθία, το οποίο χρησιμοποιεί η εν λόγω μέθοδος όσον αφορά την επιλογή σε μια συμβολοσειρά της νέας θέσης έναρξης του μοτίβου.

Έτσι, ο αλγόριθμος Gibbs Sampling για το προς εξέταση πρόβλημα έχει ως εξής:

Πίνακας 2.4: Ο Αλγόριθμος Gibbs Sampling για το Πρόβλημα.

- Επίλεξε τυχαίες θέσεις έναρξης για το μοτίβο σε όλες (N το πλήθος) τις συμβολοσειρές (μία θέση σε κάθε αλυσίδα).
- Υπολόγισε, με βάση τις προεπιλεγμένες θέσεις, τις αρχικές παραμέτρους του μοντέλου $\Psi^{(0)}$ (δηλαδή τον πίνακα θέσης βάρους $\theta^{(0)}$ που αντιστοιχεί στο μοτίβο και το διάνυσμα $b^{(0)}$ που αντιστοιχεί στο background).
- Στην $n-$οστή επανάληψη, εκτέλεσε τα παρακάτω βήματα :

Βήμα δειγματοληψίας :

- Πάρε τυχαία μια από τις N αλυσίδες, έστω την S_i, για κάποιο $i \in \{1,2,...,N\}$.
- Υπολόγισε τις παραμέτρους Ψ που προκύπτουν, αφαιρώντας τη θέση έναρξης του μοτίβου που αντιστοιχεί στην προεπιλεγμένη συμβολοσειρά S_i.
- Υπολόγισε για όλες τις πιθανές θέσεις έναρξης του μοτίβου στην S_i τις αντίστοιχες ποσότητες Q_j, $j \in \{1,2,...,L_{i-k+1}\}$.

Βήμα ενημέρωσης :

- Βρες την θέση έναρξης y στην S_i, για την οποία ισχύει ότι $Q_y = \max_{j \in \{1,2,...,L_{i-k+1}\}} \{Q_j\}$ και επίλεξέ την, ως τη νέα θέση μοτίβου στη συμβολοσειρά S_i.
- Υπολόγισε, με βάση το νέο προστιθέμενο μοτίβο, τις καινούριες παραμέτρους $\Psi^{(n)} = \{\theta^{(n)}, b^{(n)}\}$

έως ότου είναι $\theta^{(n)} = \theta^{(n-1)}$.

Αξίζει να παρατηρήσουμε ότι όπως και στον αλγόριθμο EM, το αποτέλεσμα που προκύπτει μετά την εφαρμογή του Gibbs Sampling είναι άμεσα εξαρτημένο από τις αρχικές παραμέτρους και στη δεδομένη περίπτωση, από την αρχική τυχαία επιλογή των θέσεων σε κάθε αλυσίδα, τις οποίες θεωρούμε ότι αποτελούν τις θέσεις όπου εμφανίζεται το μοτίβο.

Για να γίνει πιο κατανοητός ο παραπάνω ισχυρισμός, θεωρούμε ένα σύνολο συμβολοσειρών $S = \{S_1, S_2, S_3, S_4, S_5\}$. Υποθέτουμε ότι το βέλτιστο μοτίβο ξεκινάει σε κάθε αλυσίδα στις θέσεις 15, 1, 9, 23 και 17, αντίστοιχα, σύμφωνα με τη σειρά αρίθμησής τους. Έστω, τώρα, ότι ο αλγόριθμος έχει επιλέξει σε κάποιο αρχικό βήμα τις θέσεις έναρξης $d_1 = 17,\ \ d_2 = 3,\ \ d_3 = 11$. Μια τέτοια επιλογή είναι πολύ πιθανό να οδηγήσει σε μια μετακίνηση δύο θέσεων προς τα δεξιά από το βέλτιστο μοτίβο, γεγονός που σχετίζεται άμεσα με την αρχική επιλογή των παραμέτρων.

Παράδειγμα (εφαρμογής του Gibbs Sampling): Έστω ότι έχουμε το ακόλουθο σύνολο τεσσάρων (4) συμβολοσειρών (με σύμβολα από το αλφάβητο $\mathrm{A} = \{A, G, C, T\}$)

$$S = \{AGTTAGGCACT, GAACTTAGGC, CGTACCGCAAGCC, TTACGGAATC\}$$

και θέλουμε να προσδιορίσουμε ένα μοτίβο μήκους 4 μέσα σε αυτές.

Στο επόμενο σχήμα, γίνονται εμφανείς όλες τις δυνατές θέσεις έναρξης του μοτίβου σε κάθε αλυσίδα και συγκεκριμένα, είναι όλες εκείνες μέχρι το σύμβολο '|'.

	1 4 **8**	
S_1 :	*AGTTAGGC* | *ACT*	
	1 4 **7**	
S_2 :	*GAACTTA* | *GGC*	
	1 5 **10**	
S_3 :	*CGTACCGCAA* | *GCC*	
	1 4 **6**	
S_4 :	*TTACGG* | *AATC*	

Σχήμα 2.2: Δυνατές Θέσεις Έναρξης ενός Μοτίβου Μήκους 4 μέσα στις δοθείσες Συμβολοσειρές.

Αρχικά, ο Gibbs Sampling διαλέγει κάποια τυχαία θέση έναρξης του μοτίβου σε κάθε συμβολοσειρά, έστω τις θέσεις 5, 2, 4, 1 κατά σειρά αρίθμησης της κάθε αλυσίδας,

αντίστοιχα. Τότε, οι υπακολουθίες μήκους 4 που επιλέχθηκαν ως μοτίβο για κάθε αλυσίδα, καθώς και εκείνες που ανήκουν στο background φαίνονται στους επόμενους δύο πίνακες:

Πίνακας 2.5: Μοτίβο Αρχικής Τυχαίας Επιλογής για κάθε Αλυσίδα από τον Gibbs Sampling.

Αλυσίδα	Μοτίβο
S_1	*AGGC*
S_2	*AACT*
S_3	*ACCG*
S_4	*TTAC*

Πίνακας 2.6: Οι Υπακολουθίες που Αποτελούν το Background στο Πρόβλημα του Παραδείγματος.

AGTT	*GTTA*	*TTAG*	*TAGG*	*GGCA*	*GCAC*	*CACT*	*GAAC*	*ACTT*
CTTA	*TTAG*	*TAGG*	*AGGC*	*CGTA*	*GTAC*	*TACC*	*CCGC*	*CGCA*
GCAA	*CAAG*	*AAGC*	*AGCC*	*TACG*	*ACGG*	*CGGA*	*GGAA*	*GAAT*

Με βάση τον Πίνακα 2.5, ο πίνακας θέσης βάρους που περιγράφει το μοτίβο, είναι αρχικά

$$\theta = \begin{matrix} A \\ G \\ C \\ T \end{matrix} \begin{bmatrix} 3/4 & 1/4 & 1/4 & 0 \\ 0 & 1/4 & 1/4 & 1/4 \\ 0 & 1/4 & 1/2 & 1/2 \\ 1/4 & 1/4 & 0 & 1/4 \end{bmatrix}$$

και το διάνυσμα που αντιστοιχεί στο background (όπως προκύπτει από τον Πίνακα 2.6) είναι

$$\begin{matrix} \text{A} & \text{G} & \text{C} & \text{T} \end{matrix}$$
$$b = \begin{bmatrix} 0.2593 & 0.2963 & 0.2222 & 0.2222 \end{bmatrix}^T .$$

Ας υποθέσουμε, τώρα, ότι η συμβολοσειρά που επιλέγει ο Gibbs Sampling στο Predictive Update – Βήμα είναι η S_4. Τότε, εξαιρούμε το μοτίβο που της αντιστοιχούσε μέχρι πρότινος ($TTAC$) και οι νέες παράμετροι είναι (επιλέγοντας ως $r_A = 1,\ r_G = 1,\ r_C = 1,\ r_T = 1$, οπότε $R = 4$), σύμφωνα με τις σχέσεις (2.37) και (2.38), αντίστοιχα, προκύπτει

$$\theta = \begin{matrix} A \\ G \\ C \\ T \end{matrix} \begin{bmatrix} 4/7 & 2/7 & 1/7 & 1/7 \\ 1/7 & 2/7 & 2/7 & 2/7 \\ 1/7 & 2/7 & 3/7 & 2/7 \\ 1/7 & 1/7 & 1/7 & 2/7 \end{bmatrix}$$

και

$$b = \begin{bmatrix} 7/33 & 9/33 & 8/33 & 9/33 \end{bmatrix}^T .$$

Με βάση αυτές, υπολογίζονται τα βάρη και για τις 6 υπακολουθίες μήκους 4 που περιέχει η αλυσίδα S_4 και στη συνέχεια (βήμα Δειγματοληψίας), επιλέγεται ως νέο μοτίβο για την S_4, η υπακολουθία της με το μεγαλύτερο βάρος $Q = \max_{i=1,\ldots,6} Q_i$. Όπως και προηγουμένως, προσθέτοντας, όμως, τη νέα υπακολουθία που θεωρείται ως μοτίβο στο σύνολο των υπακολουθιών που θεωρούμε ότι αποτελούν μοτίβο και αφαιρώντας την από το background, αναθεωρούνται και πάλι οι παράμετροι θ και b. Τέλος, όπως έχουμε αναφέρει και κατά την περιγραφή του αλγορίθμου, τα δύο παραπάνω βήματα επαναλαμβάνονται έως ότου συγκλίνει ο πίνακας θ.

2.6. Τα Πιθανοτικά Δέντρα Προθεμάτων

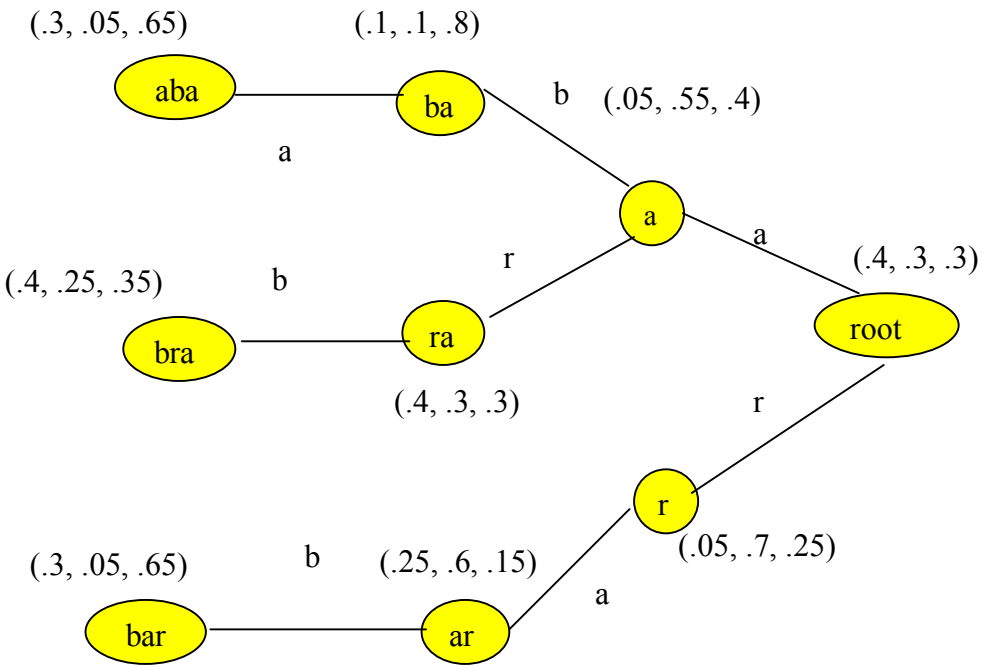
Όπως έχουμε ήδη αναφέρει, η απλούστερη κατηγορία μεθόδων που χρησιμοποιείται για τον προσδιορισμό του πιθανότερου μοτίβου μέσα σε ένα σύνολο αλυσίδων είναι η λεγόμενη απαριθμητική εξαντλητική μέθοδος. Η μέθοδος αυτή εντοπίζει όλα τα δυνατά μοτίβα καθορισμένου μήκους μέσα στις συμβολοσειρές και μετρώντας τη συχνότητα εμφάνισής τους, εκχωρεί ένα μέτρο βαρύτητας (score) ή στατιστικής σημασίας στο καθένα. Με βάση αυτό, το πιθανότερο μοτίβο είναι ασφαλώς εκείνο με το μεγαλύτερο score. Βέβαια, η μέθοδος αυτή, παρότι βρίσκει πάντα το επιθυμητό μοτίβο, είναι ασύμφορη από άποψη χρονικού κόστους, όταν το μοτίβο που αναζητούμε είναι μεγάλου μήκους ή και αρκετά περίπλοκο, μέσα σε ένα μεγάλο πάντα σύνολο δεδομένων, καθώς ο χρόνος υπολογισμού αυξάνει εκθετικά με το μήκος του μοτίβου.

Γι' αυτόν τον λόγο, αναπτύχθηκαν μια σειρά από μέθοδοι, οι οποίες χρησιμοποιούν κάποιο είδος «κλαδέματος» (δηλ. κάποιον στατιστικό περιορισμό), ώστε να μειωθεί το πλήθος των εξεταζόμενων μοτίβων [3]. Συγκεκριμένα, τέτοιου είδους μέθοδοι, όταν ψάχνουν για κάποιο μοτίβο μεγάλου μήκους, ξεκινούν την αναζήτηση από μικρότερου μήκους μοτίβα και συνεχίζουν, επεκτείνοντας μόνο εκείνα που πληρούν τις εκάστοτε συνθήκες μέχρι να φτάσουν στο ζητούμενο μήκος. Η εν λόγω διαδικασία επέκτασης του μοτίβου αναπαρίσταται με τη μορφή ενός δέντρου και μάλιστα, κάθε φορά που αυξάνουμε το μήκους του μοτίβου (κατά ένα σύμβολο), κλαδεύονται τα φύλλα εκείνα που δεν ικανοποιούν τους απαιτούμενους περιορισμούς.

Σε αυτή την κατηγορία μεθόδων, ανήκουν τα λεγόμενα Πιθανοτικά Δέντρα Προθεμάτων (Probabilistic Suffix Trees ή PSTs), τα οποία παρουσιάζουν μεγάλο ενδιαφέρον, καθώς αποτελούν μια μορφή αλυσίδας Markov μεταβλητής τάξεως (variable order Markov chain) [13]. Στη συνέχεια, ακολουθεί η περιγραφή τους και ο τρόπος εφαρμογής τους στο πρόβλημα που εξετάζουμε.

2.6.1. Γενική Περιγραφή των PSTs

Ένα πιθανοτικό δέντρο προθεμάτων (PST) που σχετίζεται με ένα συγκεκριμένο αλφάβητο $A = \{a_1, a_2, ..., a_M\}$ είναι ένα μη κενό δέντρο, του οποίου οι κόμβοι έχουν διαφορετικό βαθμό (από 0 μέχρι $|A|$ και συγκεκριμένα, τα φύλλα του έχουν βαθμό 0, ενώ στους ενδιάμεσους κόμβους ο βαθμός ποικίλλει από 1 έως $|A|$) έτσι, ώστε από κάθε κόμβο να μην εξάγονται ακμές που να αντιστοιχούν στο ίδιο γράμμα. Κάθε κόμβος αναπαριστά μια συγκεκριμένη ακολουθία συμβόλων s (λέξη) και σχετίζεται με ένα διάνυσμα πιθανοτήτων $p_s = [p_{s,1} \quad p_{s,2} \quad ... \quad p_{s,|A|}]$ μεγέθους $|A|$, το οποίο αφορά την πιθανότητα επέκτασης της λέξης (ή ακολουθίας χαρακτήρων) s που περιγράφει τον εκάστοτε κόμβο ως προς όλα τα γράμματα του αλφάβητου A και που μάλιστα, πληροί τη συνθήκη $\sum_{i=1}^{|A|} p_{s,i} = 1$. Για να γίνει πιο κατανοητός ο παραπάνω ορισμός, ακολουθεί η σχηματική αναπαράσταση ενός PST.



Σχήμα 2.3: Παράδειγμα ενός PST με βάση το αλφάβητο $A = \{a, b, r\}$.

Στο Σχήμα 2.3, απεικονίζεται ένα παράδειγμα ενός PST, το οποίο έχει δημιουργηθεί από το αλφάβητο $A = \{a, b, r\}$. Όπως παρατηρούμε, η ρίζα του δέντρου είναι ο δεξιότερος κόμβος, ενώ τα διανύσματα πιθανοτήτων που αφορούν το επόμενο σύμβολο εμφανίζονται κοντά στον αντίστοιχο κόμβο. Αξίζει να σημειώσουμε ότι τα διανύσματα αυτά σχετίζονται με την επέκταση του εκάστοτε κόμβου, προσθέτοντας κάποιο σύμβολο του αλφάβητου στα δεξιά του και μάλιστα, κάθε στοιχείο του διανύσματος αφορά το γράμμα που εμφανίζεται με την ίδια σειρά στο αλφάβητο. Έτσι, είναι φανερό ότι η κατανομή πιθανοτήτων για το επόμενο γράμμα που αφορά την υπακολουθία (χαρακτήρων) ba είναι .1, .1, .8 για τα γράμματα a, b, r, αντίστοιχα, και άρα, η πιθανότητα να εμφανιστεί το r μετά την υπακολουθία ba είναι σαφώς μεγαλύτερη για μια υπακολουθία χαρακτήρων που το μεγαλύτερο πρόθεμά της μέσα στο PST είναι το ba. Επίσης, για την αποφυγή οποιασδήποτε σύγχυσης, οφείλουμε να διαχωρίσουμε το PST από το κλασικό δέντρο προθεμάτων. Για παράδειγμα, αν θεωρήσουμε τον κόμβο (bar) στο PST του σχήματος, όπως είναι φανερό ο πατέρας του είναι ο κόμβος (ar). Αντίθετα, σε ένα κλασικό δέντρο προθεμάτων, πατέρας του ίδιου κόμβου θα ήταν ο κόμβος (ba). Το κοινό στοιχείο που παρουσιάζουν έγκειται στο ότι το PST που αντιστοιχεί σε μια καθορισμένη λέξη αποτελεί ένα υποδέντρο του κλασικού δέντρου προθεμάτων που αφορά την αντίστροφη λέξη.

2.6.2. Βασικοί Ορισμοί

Πριν προχωρήσουμε στην περιγραφή του αλγορίθμου Build – PST, στον οποίο βασίζεται η κατασκευή ενός Πιθανοτικού Δέντρου Προθεμάτων, κρίνεται απαραίτητο να ορίσουμε κάποιες έννοιες που θα συμβάλουν στην καλύτερη κατανόησή του.

Έστω A το αλφάβητο χαρακτήρων που διαθέτουμε και $r^1, r^2, ..., r^N$ ένα σύνολο από N συμβολοσειρές (αλυσίδες) με γράμματα από το A, όπου κάθε αλυσίδα r^i έχει μήκος $L_i > 0,\ i \in \{1, ..., N\}$ (και συμβολίζουμε $r^i = r_1^i r_2^i ... r_{L_i}^i$, όπου $r_j^i \in$ A).

Η *εμπειρική πιθανότητα* μιας υπακολουθίας χαρακτήρων s με βάση ένα σύνολο συμβολοσειρών $R = \{r^1, r^2, ..., r^N\}$ ορίζεται ως ο λόγος του πλήθους των εμφανίσεων της s μέσα στο R προς το πλήθος των εμφανίσεων ενός οποιουδήποτε πιθανού μοτίβου του ίδιου μήκους με την s (λαμβάνοντας υπόψη ακόμη και τις επικαλυπτόμενες εμφανίσεις ενός μοτίβου ίδιου μήκους με την s). Συγκεκριμένα, δοθείσης μιας ακολουθίας χαρακτήρων $s = s_1 s_2 ... s_k$ μήκους k και ενός συνόλου συμβολοσειρών $R = \{r^1, r^2, ..., r^N\}$, θεωρούμε ένα σύνολο μεταβλητών

$$X_s^{i,j} = \begin{cases} 1, & \text{áí } s_1 s_2 ... s_k = r_j^i r_{j+1}^i ... r_{j+k-1}^i \\ 0, & \text{äéáöïñåôéêά} \end{cases},$$

για κάθε $i = 1, ..., N,\ j = 1, ..., L_i - (k-1)$.

Όπως παρατηρούμε, η παραπάνω δείκτρια συνάρτηση παίρνει την τιμή 1, εάν η υπακολουθία s εμφανίζεται στην $j-$στή θέση της $i-$στής συμβολοσειράς του δοθέντος συνόλου, διαφορετικά λαμβάνει την τιμή 0. Τότε, η s εμφανίζεται X_s φορές μέσα στο σύνολο R, όπου

$$X_s = \sum_{i,j} X_s^{i,j}$$

και το συνολικό πλήθος των (επικαλυπτόμενων) υπακολουθιών μήκους $|s| = k$ μέσα στο ίδιο σύνολο είναι ίσο με

$$N_{|s|} = \sum_{i:L_i \geq k} [L_i - (k-1)] .$$

Έτσι, ορίζουμε ως *εμπειρική πιθανότητα* της παρατηρούμενης υπακολουθίας s τον λόγο

$$\tilde{P}(s) = \frac{X_s}{N_{|s|}} . \tag{2.39}$$

Ασφαλώς, ο τελευταίος ορισμός βασίζεται στην υπόθεση ότι οι υπακολουθίες που λαμβάνονται υπόψη στους πιο πάνω υπολογισμούς είναι ανεξάρτητες μεταξύ τους. Άρα, σύμφωνα με αυτόν τον ισχυρισμό, προκύπτει προφανώς η σχέση

$$\sum_{s \in A^k} \tilde{P}(s) = 1 . \tag{2.40}$$

Συμβολίζοντας, τώρα, με X_{s*} το πλήθος των εμφανίσεων της υπακολουθίας $s*$ στο σύνολο R, όπου $*$ είναι οποιοσδήποτε χαρακτήρας από το αλφάβητο A, δηλαδή $X_{s*} = \sum_{\sigma \in A} X_{s\sigma}$, η *δεσμευμένη εμπειρική πιθανότητα* της εμφάνισης του χαρακτήρα σ ακριβώς μετά την υπακολουθία s ορίζεται ως

$$\tilde{P}(\sigma \mid s) = \frac{X_{s\sigma}}{X_{s*}} . \tag{2.41}$$

Τέλος, συμβολίζουμε με $suf(s) = s_2 ... s_k$ (για την υπακολουθία $s = s_1 s_2 ... s_k$) και με $s^R = s_k ... s_2 s_1$ (την υπακολουθία που προκύπτει αντιστρέφοντας τους χαρακτήρες της s).

2.6.3. Ο Αλγόριθμος Build-PST

Θεωρούμε, αρχικά, ότι το μέγεθος της μνήμης για μια λέξη μέσα στο PST που κατασκευάζεται από τους χαρακτήρες ενός δοθέντος αλφάβητου A είναι ίσο με L (δηλαδή L είναι το μέγιστο μήκος μιας λέξης μέσα σε αυτό). Η διαδικασία κατασκευής ενός πιθανοτικού δέντρου προθεμάτων (PST) προχωράει σταδιακά, ξεκινώντας από όλες τις υπακολουθίες μήκους 1 (δηλαδή εκείνες που αποτελούνται από ένα γράμμα του A). Σε κάθε βήμα, το μήκος των υπακολουθιών μέσα στο δέντρο αυξάνει κατά έναν μόνο χαρακτήρα, έως ότου φθάσουμε

στο επιθυμητό (μέγιστο) μήκος L. Οι υποψήφιες υπακολουθίες, όμως, που παραμένουν στο PST (σε κάθε βήμα), πρέπει απαραιτήτως να πληρούν κάποιους συγκεκριμένους περιορισμούς.

Πιο αναλυτικά, ο αλγόριθμος Build – PST δέχεται ως είσοδο τις πέντε παραμέτρους $L, P_{\min}, r, \gamma_{\min}, \alpha,$ οι οποίες στοχεύουν στη μη εκθετική επέκταση του χώρου αναζήτησης του μοτίβου που θέλουμε να ανιχνεύσουμε. Αρχικά, το δέντρο αποτελείται από έναν μόνο κόμβο και συγκεκριμένα τη ρίζα του δέντρου, και θεωρούμε ένα σύνολο, το οποίο αποτελείται από τα γράμματα εκείνα του αλφάβητου που έχουν εμπειρική πιθανότητα μεγαλύτερη ή ίση από ένα κάτω φράγμα $P_{\min}$. Σταδιακά, σχηματίζονται όλες οι δυνατές υποψήφιες υπακολουθίες (με ένα γράμμα επιπλέον από αυτές του προηγούμενου βήματος) που μπορεί να προστεθούν στο PST και για κάθε μία από αυτές, εξετάζουμε αν η εμπειρική τους πιθανότητα είναι μεγαλύτερη ή ίση από ένα κατώτατο όριο $(1+\alpha)\cdot\gamma_{\min}$ και επιπλέον, εάν είναι αρκετά διαφορετική από την εμπειρική πιθανότητα του κόμβου που αποτελεί τον πατέρα της εκάστοτε υποψήφιας υπακολουθίας (δηλ. της συμβολοσειράς που προκύπτει αφαιρώντας το αριστερότερο γράμμα της υποψήφιας υπακολουθίας). Στην περίπτωση που ικανοποιούνται οι δύο αυτές συνθήκες, προστίθενται στο PST τόσο η υπακολουθία όσο και όλοι οι απαραίτητοι κόμβοι έτσι, ώστε να υπάρχει ένα μονοπάτι από τη ρίζα προς αυτήν.

Στο σημείο αυτό, αξίζει να αναφέρουμε ότι ένα υποψήφιο φύλλο (δηλ. μια υποψήφια υπακολουθία χαρακτήρων) σε κάποιο βήμα του αλγορίθμου θεωρείται «άχρηστο» και ασφαλώς, δεν προστίθενται στο δέντρο, εάν η εμπειρική του πιθανότητα είναι σχεδόν ίδια με αυτή του κόμβου που είναι ο πατέρας του. Παρά το γεγονός αυτό, δεν αποκλείεται η ύπαρξη τέτοιων διαδοχικών εσωτερικών κόμβων μέσα στο δέντρο, κάτι που ενδέχεται να προκύψει από την πρόσθεση ενός μοτίβου μεγαλύτερου μήκους από αυτούς και όταν οι δύο πρώτοι βρίσκονται στο μονοπάτι αυτού από τη ρίζα (από τον βαθύτερο κόμβο που υπάρχει μέχρι τη δεδομένη στιγμή και μπορεί να οδηγήσει προς το νέο μοτίβο), οπότε και κρίνεται αναγκαία και η δική τους ύπαρξη στο PST. Από τη στιγμή, λοιπόν, που προστίθεται μια νέα υπακολουθία στο δέντρο, πρέπει να υπολογισθούν και οι συναρτήσεις πρόβλεψης που αφορούν την πιθανή πρόσθεση οποιουδήποτε χαρακτήρα από το αλφάβητο στα δεξιά αυτής (δηλ. το διάνυσμα πρόβλεψης πιθανοτήτων που αντιστοιχεί σε κάθε προστιθέμενο κόμβο). Για να πάρουν την τελική τιμή τους οι συναρτήσεις αυτές, ακολουθείται μια συγκεκριμένη τεχνική έτσι, ώστε η προβλεπόμενη συνάρτηση πιθανότητας για ένα οποιοδήποτε γράμμα του

A να μην ισούται με μηδέν, αλλά στη χειρότερη περίπτωση να λαμβάνει την ελάχιστη τιμή $\gamma_{\min}$. Αυτό επιτυγχάνεται μειώνοντας τις εμπειρικές πιθανότητες έτσι, ώστε η ποσότητα $|\mathrm{A}| \cdot \gamma_{\min}$ να μοιράζεται εκ των υστέρων από όλα τα σύμβολα του A. Μάλιστα, η μείωση κάθε εμπειρικής πιθανότητας πραγματοποιείται ανάλογα με την τιμή της και συγκεκριμένα, προκύπτει από την επίλυση του συστήματος

$$\left.\begin{array}{l} \forall \sigma \in \mathrm{A} \quad \tilde{\gamma}_s(\sigma) = \lambda \cdot \tilde{P}(\sigma \mid s) + \gamma_{\min} \\ \sum_{\sigma \in \mathrm{A}} \tilde{\gamma}_s(\sigma) = 1 \end{array}\right\} \Rightarrow \sum_{\sigma \in \mathrm{A}} [\lambda \cdot \tilde{P}(\sigma \mid s) + \gamma_{\min}] = 1$$

$$\Rightarrow \quad \lambda \cdot 1 + \gamma_{\min} |\mathrm{A}| = 1 \quad \Rightarrow \quad \lambda = 1 - \gamma_{\min} |\mathrm{A}| .$$

Απαιτώντας, τώρα, να ισχύει $\lambda \geq 0 \Rightarrow \gamma_{\min} \leq \frac{1}{|\mathrm{A}|}$, οπότε προκύπτει

$$\tilde{\gamma}_s(\sigma) = \left(1 - \gamma_{\min} \cdot |\mathrm{A}|\right) \cdot \tilde{P}(\sigma \mid s) + \gamma_{\min} .$$

Από την τελευταία σχέση, είναι προφανές ότι όταν είναι $\tilde{P}(\sigma \mid s) = 0,$ ισχύει $\tilde{\gamma}_s(\sigma) = \gamma_{\min}$.

Συμβολίζοντας, στη συνέχεια, με T το πιθανοτικό δέντρο προθεμάτων και με S το σύνολο των συμβολοσειρών που θέλουμε κάθε φορά να εξετάσουμε, ο αλγόριθμος Build – PST υπό τη μορφή ψευδοκώδικα περιγράφεται ως εξής:

Πίνακας 2.7: Ο Αλγόριθμος Build – PST για την κατασκευή ενός Πιθανοτικού Δέντρου Προθεμάτων.

- Αρχικά, το δέντρο αποτελείται από τη ρίζα και θεωρούμε το σύνολο $S = \left\{ \sigma \mid \sigma \in \mathrm{A} : \quad \tilde{P}(\sigma) \geq P_{\min} \right\}$.
- Κατασκευή του PST :

Όσο ισχύει ότι $S \neq \varnothing$, για κάθε στοιχείο $s \in S$ εκτέλεσε τα ακόλουθα βήματα

- Αφαίρεσε το s από το S.
- Αν υπάρχει κάποιο σύμβολο $\sigma \in \mathrm{A}$ τέτοιο, ώστε

$$\tilde{P}(\sigma \mid s) \geq (1+\alpha) \cdot \gamma_{\min}$$

και

$$\frac{\tilde{P}(\sigma \mid s)}{\tilde{P}(\sigma \mid suf(s))} \begin{cases} \geq r \\ \eta \\ \leq 1/r \end{cases},$$

τότε πρόσθεσε στο T τον κόμβο που αντιστοιχεί στο s και όλους τους κόμβους για το μονοπάτι προς το s από τον βαθύτερο κόμβο μέσα στο T που είναι πρόθεμα του s.

- Εάν ισχύει ότι $|s| < L$, τότε πρόσθεσε όλες τις υπακολουθίες $\left\{ \sigma' s \mid \sigma' \in \mathrm{A} \quad \kappa\alpha\iota \quad \tilde{P}(\sigma' s) \geq P_{\min} \right\}$ (αν υπάρχουν) στο S.
- Για κάθε s που αποτελεί έναν κόμβο στο T, υπολόγισε την $\tilde{\gamma}_s(\sigma) = \left(1 - \gamma_{\min} \cdot |\mathrm{A}|\right) \cdot \tilde{P}(\sigma \mid s) + \gamma_{\min}$.

Εφόσον έχει πια προηγηθεί η αναλυτική περιγραφή του αλγορίθμου, μπορούμε να ανατρέξουμε στο Σχήμα 2.3 και να υπολογίσουμε αναλυτικά την πιθανότητα πρόβλεψης (με βάση το δοθέν PST) μιας οποιασδήποτε συμβολοσειράς s με χαρακτήρες από το αλφάβητο $A = \{a, b, r\}$. Αν, λοιπόν, θεωρήσουμε τη συμβολοσειρά "*barbara*", τότε έχουμε

$$\tilde{P}(barbara) = \tilde{P}(b) \cdot \tilde{P}(a \mid \underline{b}) \cdot \tilde{P}(r \mid \underline{ba}) \cdot \tilde{P}(b \mid \underline{bar}) \cdot \tilde{P}(a \mid barb) \cdot \tilde{P}(r \mid bar\underline{ba}) \cdot \tilde{P}(a \mid bar\underline{bar})$$

$$= \tilde{\gamma}_{root}(b) \cdot \tilde{\gamma}_{b}(a) \cdot \tilde{\gamma}_{ba}(r) \cdot \tilde{\gamma}_{bar}(b) \cdot \tilde{\gamma}_{root}(a) \cdot \tilde{\gamma}_{ba}(r) \cdot \tilde{\gamma}_{bar}(a)$$

$$= 0.3 \cdot 0.2 \cdot 0.8 \cdot 0.05 \cdot 0.4 \cdot 0.8 \cdot 0.65 = 4.992 \cdot 10^{-4}.$$

Για να κατανοήσει κανείς πώς υπολογίσθηκε η παραπάνω πιθανότητα πρόβλεψης, θα πρέπει να διαχωρίσει δύο πράγματα που αφορούν ένα PST. Συγκεκριμένα, παρ' ότι ένας κόμβος που προσθέτουμε στο δέντρο έχει ένα ακόμα γράμμα στο αριστερότερο άκρο του σε σχέση με τον κόμβο που αποτελεί τον πατέρα του, οι πιθανότητες πρόβλεψης αφορούν την πρόσθεση ενός επιπλέον χαρακτήρα στο δεξιότερο άκρο της εκάστοτε λέξης. Έτσι, η υπογραμμισμένη ακολουθία χαρακτήρων, στον υπολογισμό της πιθανότητας πρόβλεψης για τη συμβολοσειρά "*barbara*", αντιστοιχεί στο μεγαλύτερο δυνατό πρόθεμα που εμφανίζεται στο PST (και βρίσκεται δεξιά του '|'), ενώ η αντίστοιχη πιθανότητα που εμφανίζεται στον υπολογισμό αποτελεί τη συνάρτηση πρόβλεψης που αφορά τον επόμενο χαρακτήρα (ο οποίος βρίσκεται στο αριστερό μέρος του '|') δεδομένου του μεγαλύτερου δυνατού υπαρκτού προθέματος μέσα στο PST. Επομένως, το βήμα πρόβλεψης χρησιμοποιείται για την πρόσθεση ενός χαρακτήρα στο δεξιό τμήμα μιας συμβολοσειράς, δοθέντος του μεγαλύτερου προθέματός της μέσα στο δέντρο, ξεκινώντας την εξέταση της συμβολοσειράς από αριστερά προς τα δεξιά.

Η απόδοση του αλγόριθμου που περιγράψαμε παρουσιάζει σαφές πλεονέκτημα, συγκρινόμενη με τις περισσότερες μεθόδους που έχουν κατασκευαστεί για τον ίδιο σκοπό όπως είναι ο EM και ο Gibbs Sampling, διότι δεν εξαρτάται από την αρχικοποίηση κάποιων παραμέτρων και δεν απαιτεί ευθυγράμμιση των συμβολοσειρών που συνιστούν το σύνολο δεδομένων του προβλήματος, ενώ είναι εξίσου καλή με την απόδοση ενός κρυμμένου μοντέλου Markov για ολική αναζήτηση σημαντικών μοτίβων.

Αξίζει να σημειώσουμε, βασιζόμενοι στο τελευταίο παράδειγμα, ότι ένα PST θα μπορούσε να χαρακτηριστεί ως ένα μοντέλο που βασίζεται σε μεταβλητής τάξης (εδώ, μήκους του εκάστοτε μεγαλύτερου δυνατού προθέματος μέσα στο PST) αλυσίδες Markov, διότι ενώ στην περίπτωση των αλυσίδων Markov η πιθανότητα ενός ενδεχομένου εξαρτάται ακριβώς από τις k το πλήθος προηγούμενες καταστάσεις, σε ένα πιθανοτικό δέντρο πιθανοτήτων (όπως δείξαμε στο προηγούμενο παράδειγμα) η εξάρτηση αυτή στηρίζεται σε ένα διαφορετικό κάθε φορά πλήθος προηγούμενων γεγονότων που ποικίλλει από 0 έως k ανάλογα με το μήκος του μεγαλύτερου δυνατού προθέματος της ακολουθίας χαρακτήρων μέσα στο PST, της οποίας θέλουμε να προβλέψουμε την πιθανότητα. Συνεπώς, τα πιθανοτικά δέντρα προθεμάτων αποτελούν μια γενίκευση των αλυσίδων Markov.

Τέλος, ο αλγόριθμος που περιγράψαμε προτείνει μια πολύ απλή λύση σε ένα πρόβλημα ιδιαίτερα σημαντικό για μοντέλα βασισμένα σε αλυσίδες Markov, το οποίο αφορά την «εξομάλυνση» της προβλεπόμενης πιθανότητας μιας υπακολουθίας. Το πρόβλημα αυτό παρουσιάζεται συχνά στην περίπτωση που το πλήθος των πιθανών ενδεχομένων είναι πολύ μεγάλο σε σχέση με το πλήθος αυτών των παρατηρούνται. Είναι πολύ πιθανό, λοιπόν, τότε να εκτιμηθεί για πολλά μη παρατηρούμενα (πιθανά) ενδεχόμενα ότι έχουν μηδενική πιθανότητα να εμφανιστούν. Η διαδικασία που ακολουθείται για την εξομάλυνση των προβλεπόμενων πιθανοτήτων κατορθώνει να μην προβλέπει μηδενική πιθανότητα εμφάνισης τέτοιων ενδεχομένων (αλλά να της αποδίδει ως ελάχιστη τιμή ίση με $\gamma_{\min}$) και συγχρόνως, να διορθώνει την τιμή εκείνων που εμφανίζονται μέσα στο σύνολο εκπαίδευσης.

ΚΕΦΑΛΑΙΟ 3. ΤΕΧΝΙΚΕΣ ΑΡΧΙΚΟΠΟΙΗΣΗΣ ΤΩΝ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΜΟΤΙΒΟΥ

Όπως φαίνεται από τις μεθόδους που εξετάσαμε προηγουμένως, ένα σημαντικό ζήτημα που προκύπτει είναι η αρχικοποίηση των παραμέτρων τους και κατά βάση, του πίνακα θ που παριστάνει το μοτίβο. Το πρόβλημα αυτό, το οποίο παρουσιάζεται στον Gibbs Sampling, αλλά κυρίως στον EM, δημιούργησε την ανάγκη περαιτέρω έρευνας με σκοπό την αναζήτηση μιας καλής αρχικοποίησης για την παράμετρο που μας ενδιαφέρει. Μια λύση που προτάθηκε, ήταν αυτή της ομαδοποίησης των παρατηρούμενων δεδομένων, η οποία συνήθως δίνει μια καλή αρχικοποίηση για τον θ. Συγκεκριμένα, η ιδέα αυτή προέκυψε από την πεποίθηση ότι τα δείγματα που αναπαριστούν το μοτίβο παρουσιάζουν πολλά κοινά χαρακτηριστικά μεταξύ τους και πιθανότατα, κατόπιν ομαδοποίησης όλου του συνόλου των δεδομένων, να συνιστούν την ομάδα εκείνη που τα στοιχεία της διαφέρουν λιγότερο μεταξύ τους σε σύγκριση με αυτά των υπολοίπων ομάδων που σχηματίζονται. Συνεπώς, μια αποτελεσματική τεχνική ομαδοποίηση, κάλλιστα θα μπορέσει να μας παρέχει και μια καλή αρχικοποίηση του πίνακα που περιγράφει το μοτίβο.

Στις ενότητες που ακολουθούν, θα εξετάσουμε δύο πολύ γνωστές μεθόδους ομαδοποίησης δεδομένων, καθώς και τον τρόπο με τον οποίο τις προσαρμόσαμε στο πρόβλημα που

μελετούμε. Μάλιστα, στο 5^{ο} κεφάλαιο, θα τις χρησιμοποιήσουμε και πειραματικά για τον προσδιορισμό μιας αρχικής εκτίμησης της βασικής παραμέτρου του προβλήματος και θα διαπιστώσουμε ότι η τελευταία είναι δυνατόν να προσαρμοστεί ακόμα καλύτερα στα πραγματικά δεδομένα του προβλήματος, εφαρμόζοντας στη συνέχεια κάποια τεχνική βελτιστοποίησης (αλγόριθμος ΕΜ (κεφ. 2), παραλλαγές του ΕΜ(κεφ.4)), όπου και θα γίνει εμφανής η χρησιμότητά τους.

3.1. Η Έννοια της Ομαδοποίησης

Με τον όρο *ομαδοποίηση* ενός συνόλου δεδομένων εννοούμε τον διαχωρισμό του σε υποσύνολα (ομάδες) έτσι, ώστε κάθε αντικείμενο να ανήκει σε ένα μόνο από αυτά τα υποσύνολα και η επιλογή αυτή να βασίζεται σε πληροφορίες που πηγάζουν μόνο από τα ίδια τα δεδομένα και τις μεταξύ τους σχέσεις.

Ο κύριος στόχος της ομαδοποίησης είναι η ύπαρξη κοινών χαρακτηριστικών μεταξύ των αντικειμένων κάθε ομάδας τέτοιων, ώστε να τα διαφοροποιούν από τα αντικείμενα των υπόλοιπων ομάδων. Γι' αυτό, όσο πιο πολύ διαφέρουν μεταξύ τους οι ομάδες που σχηματίζονται, τόσο πιο επιτυχημένη θα μπορεί να χαρακτηριστεί η ομαδοποίηση ενός συνόλου δεδομένων. Ασφαλώς, είναι εξαιρετικά δύσκολο να προσδιορίσει κανείς την έννοια της ομάδας, καθώς αυτό εξαρτάται αποκλειστικά από το εκάστοτε σύνολο δεδομένων που διατίθεται.

Τα τελευταία χρόνια έχει αναπτυχθεί ένα μεγάλο πλήθος τεχνικών που μπορούν να χρησιμοποιηθούν για τον διαχωρισμό ενός συνόλου αντικειμένων σε ομάδες. Εδώ, θα ασχοληθούμε με δύο είδη ομαδοποίησης, τη διαμεριστική (partitional) και την ιεραρχική (hierarchical) και συγκεκριμένα, με μια παραλλαγή του αλγόριθμου των Κ–κέντρων (K-means) και του Συνθετικού αλγορίθμου (Agglomerative) [6,16,19,25], οι οποίοι ανήκουν στο πρώτο και το δεύτερο είδος ομαδοποίησης, αντίστοιχα..

3.2. Ο Αλγόριθμος των Κ-κέντρων (K-means)

Μια πολύ δημοφιλής τεχνική ομαδοποίησης ενός συνόλου δεδομένων, η οποία ξεχωρίζει για την απλότητά της, είναι ο αλγόριθμος των Κ-κέντρων. Στην προκειμένη περίπτωση, μια ομάδα περιγράφεται από έναν *αντιπρόσωπο* και όλα τα σημεία που ανήκουν σε αυτήν απέχουν μικρότερη απόσταση από τον αντιπρόσωπό της σε σχέση με τους αντιπροσώπους των λοιπών ομάδων. Για παράδειγμα, στην περίπτωση των διακριτών δεδομένων (δηλ. με διακριτά χαρακτηριστικά) ως αντιπρόσωπος θα μπορούσε να θεωρηθεί το πιο αντιπροσωπευτικό στοιχείο της ομάδας.

Η μέθοδος έχει ως βασικό στόχο τη διαμέριση του συνόλου δεδομένων σε ένα καθορισμένο πλήθος Κ ομάδων. Οι ομάδες αυτές αναπαρίστανται από κάποιους αντιπροσώπους, οι οποίοι χαρακτηρίζονται ως κέντρα (ή κεντροειδή) των ομάδων και συνήθως, αποτελούν τη μέση τιμή των στοιχείων κάθε ομάδας. Από τον τρόπο, βέβαια, που ορίζονται τα κέντρα, καθίσταται προφανές το γεγονός ότι είναι σπάνιο φαινόμενο το κέντρο μιας ομάδας να αντιστοιχεί σε ένα στοιχείο του συνόλου δεδομένων.

Για να κατορθώσει, λοιπόν, ο αλγόριθμος των Κ-κέντρων να χωρίσει το σύνολο δεδομένων σε Κ το πλήθος ομάδες, ακολουθεί μια σειρά από απλά βήματα. Αρχικά, επιλέγονται Κ τυχαία κέντρα. Εν συνεχεία, κάθε σημείο του συνόλου (ή στοιχείο) ανατίθεται στο κοντινότερο κεντροειδές, χρησιμοποιώντας κάποιο μέτρο απόστασης. Με αυτόν τον τρόπο, το σύνολο των σημείων που έχουν ανατεθεί σε ένα συγκεκριμένο κεντροειδές σχηματίζουν μια ομάδα. Με βάση, τώρα, τα σημεία της κάθε ομάδας, το κέντρο της επαναπροσδιορίζεται. Τα δύο αυτά βήματα του αλγορίθμου επαναλαμβάνονται έως ότου (για πρώτη φορά) δεν μεταβληθούν τα κέντρα των ομάδων και συνεπώς, έχουν πια σχηματιστεί οι ζητούμενες Κ ομάδες. Μια συνοπτική περιγραφή του αλγορίθμου παρατίθεται στον πίνακα που ακολουθεί.

Πίνακας 3.1: Ο Αλγόριθμος των Κ-κέντρων.

- Αρχικά, διάλεξε Κ τυχαία σημεία ως τα κέντρα των ομάδων που πρόκειται να σχηματιστούν.
- Επανάλαβε τα παρακάτω βήματα:
  - Δημιούργησε Κ ομάδες, αναθέτοντας κάθε σημείο στο κοντινότερο κέντρο.

- Επαναϋπολόγισε το κέντρο κάθε ομάδας βάσει των σημείων που της έχουν ανατεθεί

έως ότου (για πρώτη φορά) δεν μεταβληθούν τα κέντρα.

Όπως μπορεί εύκολα να διακρίνει κανείς, ένα πολύ βασικό στοιχείο του παραπάνω αλγορίθμου είναι η επιλογή του μέτρου απόστασης που θα χρησιμοποιήσουμε, προκειμένου να αναθέσουμε κάθε σημείο σε μία από τις ομάδες. Για τον σκοπό αυτό, χρησιμοποιείται συνήθως η Ευκλείδεια απόσταση, που ως γνωστόν για δύο σημεία $x, y \in \Re^n$ με $x = [x_1 \quad ... \quad x_n], y = [y_1 \quad ... \quad y_n]$, ορίζεται ως

$$d(x,y) = \left(\sum_{i=1}^{n} | x_i - y_i |^2 \right)^{1/2}.$$

Παρά το γεγονός ότι ο αλγόριθμος των K-κέντρων είναι υπολογιστικά απλός, απαιτεί μεγάλο πλήθος υπολογισμών για εξαιρετικά μεγάλα σύνολα δεδομένων (για τις αποστάσεις κάθε σημείου από όλα τα κέντρα των ομάδων). Όσον αφορά την πολυπλοκότητα χρόνου που απαιτεί, αυτή ανέρχεται σε $O(NKld)$, για N σημεία με d διαστάσεις, K ομάδες και l επαναλήψεις. Επιπλέον, ένα βασικό μειονέκτημά του είναι η τυχαία αρχικοποίηση των κέντρων, διότι το αποτέλεσμα που θα δώσει έχει άμεση εξάρτηση από την αρχική επιλογή των κέντρων. Τέλος, δεν αντιλαμβάνεται την ύπαρξη *outliers*, τα οποία ενδέχεται να επηρεάσουν σε μεγάλο βαθμό τον τελικό σχηματισμό των ομάδων, με άμεση συνέπεια να παραχθεί κάποιο αποτέλεσμα ομαδοποίησης που να μην αντιστοιχεί καθόλου στην πραγματικότητα. Έτσι, ο καλύτερος τρόπος αντιμετώπισης του τελευταίου προβλήματος είναι ο εντοπισμός των outliers και η απομάκρυνσή τους πριν από την εφαρμογή του αλγορίθμου.

3.3. Μια Παραλλαγή του Αλγορίθμου των K-κέντρων

Προκειμένου να υπάρχει κάποιος έλεγχος όσον αφορά το πλήθος των ομάδων με υπακολουθίες (μήκους K) που πρόκειται να παράγει ο αλγόριθμος, τον προκαθορίσαμε ίσο με r / N, όπου N είναι το πλήθος των αρχικά διδόμενων συμβολοσειρών $S = \{S_1, ..., S_N\}$ και r το πλήθος των υπακολουθιών μήκους K που προκύπτουν από το σύνολο S.

Ο σχηματισμός, τώρα, των ομάδων ακολουθεί την ίδια διαδικασία του αλγόριθμου των Κ-κέντρων, με τη διαφορά ότι αντί της Ευκλείδειας απόστασης, χρησιμοποιούμε την απόσταση Manhattan που για δύο σημεία $x, y \in \Re^n$ με $x = [x_1 \quad ... \quad x_n], y = [y_1 \quad ... \quad y_n]$, υπολογίζεται από τη σχέση

$$d(x, y) = \sum_{i=1}^{n} | x_i - y_i | = \| x - y \|_1 .$$

Το βασικό ερώτημα, βέβαια, που μπορεί να τεθεί είναι πώς είναι δυνατόν να ορίσουμε κάποιο μέτρο απόστασης μεταξύ ακολουθιών σταθερού μήκους K, οι οποίες αποτελούνται από σύμβολα ενός δεδομένου αλφάβητου $A = \{a_1, ..., a_M\}$. Η απάντηση στο ερώτημα αυτό δίδεται άμεσα, εάν σκεφτεί κανείς ότι κάθε τέτοια υπακολουθία μπορεί να παρασταθεί υπό τη μορφή ενός MxK πίνακα βάρους θέσης, όπου στην (i, j) θέση του αντιστοιχεί η σχετική συχνότητα εμφάνισης του χαρακτήρα a_i στην $j-$στη θέση της υπακολουθίας. Προφανώς, επειδή στην περίπτωση της μίας υπακολουθίας, έστω της $X_l = [X_{l1} ... X_{lK}]$, σε κάθε θέση της υπάρχει ένα και μόνο γράμμα του αλφάβητου, συμβολίζοντας τον αντίστοιχο πίνακα με $\theta^{(X)} = [\theta_{ij}^{(X)}]$, αυτός ορίζεται ως εξής:

$$\theta_{ij}^{(X)} = \begin{cases} 1, \text{αν} \quad \mathrm{X_{lj}} = a_i \\ 0, \text{διαφορετικά} \end{cases}, \; i = 1, ..., M, \;\; j = 1, ..., K .$$

Από τον ορισμό του παραπάνω πίνακα, είναι φανερό ότι ισχύει $\sum_{i=1}^{M} \theta_{ij}^{(X)} = 1,$ για κάθε $j = 1, ..., K$.

Με βάση αυτόν τον πίνακα, αρχικά, που αντιστοιχεί σε κάθε υπακολουθία υπολογίζονται οι μεταξύ τους αποστάσεις Manhattan και συγκεκριμένα, ορίζονται ως το άθροισμα των απολύτων διαφορών των αντίστοιχων στοιχείων τους (και αυτό, γιατί θεωρούμε νοητά τον κάθε MxK πίνακα ως ένα $MKx1$ διάνυσμα). Όπως υπαγορεύει ο αλγόριθμος των Κ-κέντρων, αφού υπολογισθούν οι αποστάσεις αυτές μεταξύ των υπακολουθιών που έχουν επιλεγεί ως κέντρα και των υπολοίπων, αναθέτουμε κάθε υπακολουθία στην ομάδα από το κέντρο της οποίας απέχει τη μικρότερη απόσταση. Στη συνέχεια, υπολογίζουμε το νέο κέντρο της κάθε ομάδας ως τον μέσο όρο των στοιχείων (πινάκων) που τις ανατέθηκαν προηγουμένως.

Προφανώς, τα νέα κέντρα ενδέχεται να μην αντιστοιχούν σε κάποια υπακολουθία (δηλαδή σε κάποιο στοιχείο του συνόλου δεδομένων, όπως συμβαίνει γενικά κατά την εφαρμογή του αλγορίθμου). Παρά το γεγονός, όμως, αυτό, και πάλι περιγράφεται το καθένα από έναν αριθμητικό MxK πίνακα, έστω συμβολικά $\theta^{\mathrm{K}_l} = [\theta_{ij}^{\mathrm{K}_l}]$ ο πίνακας που αντιστοιχεί στο κέντρο της $l-$στης ομάδας, ο οποίος εξακολουθεί να ικανοποιεί τις συνθήκες $\sum_{i=1}^{M} \theta_{ij}^{\mathrm{K}_{li}} = 1,$ για κάθε $j = 1,...,K$.

Παραδείγματος χάριν, εάν θεωρήσουμε το αλφάβητο $\Sigma = \{A, B, C, E\},$ η υπακολουθία $ABBACE$ περιγράφεται από τον ακόλουθο $4x6$ αριθμητικό πίνακα

$$\theta^{(ABBACE)} = \begin{array}{c} \\ \downarrow \\ A \\ B \\ C \\ E \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & B & A & C & E \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

Υποθέτουμε, ακόμη, ότι όλο το σύνολο παρατηρήσεων που διαθέτουμε είναι το $X = \{ABBECE, CAEEBA, ACEEAA, ABBACE\}$ που περιέχει υπακολουθίες μήκους 6 από χαρακτήρες του Σ και ότι στόχος μας είναι να δημιουργήσουμε δύο ομάδες.

Έστω ότι, αρχικά, επιλέγονται ως κέντρα των δύο ομάδων οι υπακολουθίες $X_1 = ABBECE$ και $X_2 = CAEEBA$. Τότε, υπολογίζοντας την απόσταση Manhattan (L_1) των υπόλοιπων υπακολουθιών από τα κέντρα των ομάδων, έχουμε $d(X_1, X_3) = 4,$ $d(X_2, X_3) = 3,$ $d(X_1, X_4) = 1,$ $d(X_2, X_4) = 6,$ απ' όπου προκύπτει ότι $d(X_2, X_3) = \min_{i=1,2} d(X_i, X_3)$ και $d(X_1, X_4) = \min_{i=1,2} d(X_i, X_4)$. Επομένως, οι υπακολουθίες X_3 και X_4 ανατίθενται στις ομάδες με κέντρα τις X_2 (2^{η} ομάδα) και X_1 (1^{η} ομάδα), αντίστοιχα. Άρα, τα νέα κέντρα των δύο ομάδων, σύμφωνα με όσα ειπώθηκαν προηγουμένως, αναπαρίστανται από τους πίνακες (κατά σειρά για την πρώτη και τη δεύτερη ομάδα)

$$\theta^{K_1} = \frac{1}{2}\left(\theta^{(ABBACE)} + \theta^{(ABBECE)}\right) = \begin{bmatrix} 1 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 1 \end{bmatrix}$$

και

$$\theta^{K_2} = \frac{1}{2}\left(\theta^{(CAEEBA)} + \theta^{(ACEEAA)}\right) = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 & 1/2 & 1 \\ 0 & 0 & 0 & 0 & 1/2 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \end{bmatrix}.$$

Εάν υπολογίσουμε και πάλι τις αποστάσεις όλων των υπακολουθιών από τα νέα κέντρα, θα παρατηρήσουμε ότι οι ομάδες δεν αλλάζουν και συνεπώς, ούτε τα κέντρα. Αυτό σημαίνει ότι ο αλγόριθμος τερματίζει σε αυτό το σημείο. Εδώ, είναι λογικό να αναρωτηθεί κανείς τι ακριβώς κερδίσαμε με την ομαδοποίηση των δεδομένων. Αυτό που ελπίζουμε είναι να υπάρχει σε μία από αυτές τις ομάδες, η κρυμμένη πληροφορία που αναζητούμε ή έστω ένα μέρος αυτής. Συγκεκριμένα, αν υποθέσουμε ότι θέλουμε να βρούμε σε ποια ομάδα του προηγούμενου παραδείγματος «κρύβεται» το μοτίβο $m = ABBECE$, τότε προφανώς αυτό περιγράφεται καλύτερα από το κέντρο της πρώτης ομάδας.

Ασφαλώς, ένα σημαντικό θέμα που προκύπτει είναι με ποιο κριτήριο θα γίνει η επιλογή της καταλληλότερης από αυτές τις ομάδες, δηλαδή εκείνης που αποτελεί την καλύτερη δυνατή αναπαράσταση του μοτίβου. Επειδή, όπως αναφέρθηκε στην προηγούμενη παράγραφο, το αποτέλεσμα που παράγει ο αλγόριθμος των K-κέντρων εξαρτάται κατά πολύ από την τυχαία αρχικοποίηση των κέντρων, θα καταφύγουμε στη λύση της κατ' επανάληψιν εφαρμογής του στο ίδιο σύνολο δεδομένων και κατόπιν, της τελικής επιλογής του πίνακα (ή ισοδύναμα της ομάδας) που περιγράφει καλύτερα το μοτίβο (από όλες τις εφαρμογές του αλγορίθμου). Αυτό το σκεπτικό ακολουθούμε στα πειράματα του 5ου κεφαλαίου, όπου φυσικά γνωρίζουμε τον πίνακα βάρους θέσης που αναπαριστά το ζητούμενο μοτίβο (και αυτό γνωρίζοντας τις θέσεις σε κάθε αλυσίδα που αποτελούν θέσεις μοτίβου). Η γνώση αυτή, βέβαια, δεν χρησιμοποιείται κατά την επίλυση του προβλήματος, παρά μόνο για την επίτευξη μιας καλής αρχικοποίησης του εν λόγω πίνακα και όχι μιας τυχαίας, η οποία, στη γενική περίπτωση, δεν θα δώσει ικανοποιητική λύση στο πρόβλημα εντοπισμού των υπακολουθιών που περιγράφουν το μοτίβο.

Συνοπτικά, ο αλγόριθμος των K-κέντρων, για το πρόβλημα, παρατίθεται στον Πίνακα 3.2.

Πίνακας 3.2: Η Εφαρμογή του Αλγόριθμου των K-κέντρων στο Πρόβλημα.

- Επίλεξε τυχαία r/N το πλήθος υπακολουθίες από το σύνολο X ως τα αρχικά κέντρα των ομάδων.
- Υπολόγισε για κάθε $X_i \in X$ τον πίνακα βάρους θέσης $\theta^{(X_i)}$ που της αντιστοιχεί.
- Επανάλαβε τα παρακάτω βήματα:
  - Υπολόγισε, με βάση τους πίνακες για κάθε X_i, τις αποστάσεις Manhattan των υπακολουθιών από τα κέντρα των ομάδων K_{cl_j}.
  - Ανάθεσε κάθε X_i στην ομάδα, από το κέντρο της οποίας απέχει την ελάχιστη απόσταση.
  - Με βάση τις ομάδες που σχηματίστηκαν, υπολόγισε το νέο κέντρο κάθε ομάδας $(cl_j)_{j=1}^{r/N}$ από τον τύπο (δηλαδή τον πίνακα που το περιγράφει)

$$\theta^{(\mathrm{K}_{cl_j})} = \frac{\sum_{X_i \in cl_j} \theta^{(X_i)}}{\left|cl_j\right|},$$

όπου $\left|cl_j\right|$ το πλήθος των υπακολουθιών που υπάγονται στην $j-$στη ομάδα

έως ότου τα κέντρα παραμείνουν σταθερά.

3.4. Ο Συνθετικός Αλγόριθμος Ομαδοποίησης

Μια διαφορετικού είδους προσέγγιση για την ομαδοποίηση δεδομένων, σε σύγκριση με τον αλγόριθμο των K-κέντρων, χρησιμοποιεί ο λεγόμενος συνθετικός αλγόριθμος. Πρόκειται για μια τεχνική που ανήκει στην κατηγορία των ιεραρχικών μεθόδων ομαδοποίησης δεδομένων. Αυτή παράγει ένα σύνολο ομάδων που κάθε μία έχει προέλθει από τη συνένωση δύο άλλων ομάδων σε κάποιο βήμα. Μάλιστα, τέτοιου είδους μέθοδοι ομαδοποίησης μπορούν να παρασταθούν με τη μορφή ενός δενδροδιαγράμματος, στο οποίο γίνονται πιο εμφανείς οι σχέσεις αυτές μεταξύ των ομάδων (δηλαδή από τη συνένωση ποιων υποομάδων

έχει προκύψει κάθε ομάδα). Συγκεκριμένα, κάθε εσωτερικός κόμβος αποτελεί ένωση των παιδιών του, ενώ η ρίζα του δέντρου θα μπορούσε να νοηθεί ως η μεγαλύτερη δυνατή ομάδα, καθώς περιέχει όλα τα στοιχεία προς ομαδοποίηση.

Προχωρώντας στη γενική περιγραφή του αλγορίθμου, αρχικά, κάθε σημείο (αντικείμενο) του συνόλου δεδομένων αποτελεί μια ομάδα. Σε κάθε βήμα, συνενώνονται οι δύο κοντινότερες ομάδες βάσει ενός μέτρου εγγύτητας που έχουμε επιλέξει. Η διαδικασία αυτή της συνένωσης των ομάδων επαναλαμβάνεται μέχρι, τελικά, να απομείνει μία μόνο ομάδα. Η συνθετική μέθοδος ομαδοποίησης, με τη μορφή ψευδοκώδικα, έχει ως εξής:

Πίνακας 3.3: Ο Συνθετικός Αλγόριθμος Ομαδοποίησης.

- Για το σύνολο δεδομένων $S = \{s_1, s_2, ..., s_n\}$, υπολόγισε αρχικά ένα μέτρο της μεταξύ τους απόστασης $dist(s_i, s_j)$, για κάθε $i, j \in \{1,2,...,n\}$, $i \neq j$ και θεώρησε κάθε s_i ως μια ομάδα cl_i.
- Επανάλαβε τα ακόλουθα βήματα:
  - Συνένωσε τις δύο κοντινότερες ομάδες, δηλαδή τις ομάδες cl_i, cl_j, για τις οποίες ισχύει $dist(cl_i, cl_j) = \min_{\substack{k,l \\ k \neq l}} [dist(cl_k, cl_l)]$.
  - Υπολόγισε τα μέτρα $dist(cl_i, cl_j)$ για όλες τις ομάδες (ανά δύο) που υπάρχουν

  έως ότου απομείνει μία μόνο ομάδα.

Εφόσον, τώρα, οι αποστάσεις μεταξύ των ομάδων επαναϋπολογίζονται μετά από κάθε συγχώνευση ανάμεσα σε δύο ομάδες, είναι φανερό ότι το υπολογιστικό κόστος του αλγορίθμου είναι πολύ μεγάλο, με αποτέλεσμα να τον καθιστά εξαιρετικά αργό. Συγκεκριμένα, απαιτεί $O(N^2 \log N)$ χρόνο για N το πλήθος ομάδες (αντικείμενα) αρχικά και $O(N^2)$ χώρο. Παρά ταύτα, είναι ιδιαίτερα αποτελεσματικός, όσον αφορά την ομαδοποίηση δεδομένων, αφού γενικά παίρνει «σωστές» αποφάσεις συνένωσης ομάδων σε κάθε βήμα. Άμεση συνέπεια αυτού, είναι ο σχηματισμός καλής ποιότητας ομάδων (δηλαδή ομάδων που ξεχωρίζουν μεταξύ τους). Τέλος, ένα σημαντικό πλεονέκτημα της μεθόδου είναι

ότι δεν επηρεάζεται από την ύπαρξη αντικειμένων με πολύ διαφορετικά χαρακτηριστικά από τα υπόλοιπα (εμφάνιση *outliers*).

3.5. Η Εφαρμογή του Συνθετικού Αλγόριθμου στο Πρόβλημα

Με τους συμβολισμούς που έχουμε μέχρι στιγμής εισαγάγει, και στην περίπτωση του συνθετικού αλγόριθμου, δεν ακολουθούμε τα ακριβή βήματά του κατά την εφαρμογή του στο πρόβλημα που εξετάζουμε. Στην ουσία, το βήμα που αναθεωρούμε είναι αυτό της συνένωσης όλων των ομάδων (εδώ, υπακολουθιών μήκους k) σε μία τελική που περιλαμβάνει όλες τις αρχικές. Οι επαναλήψεις του αλγορίθμου ολοκληρώνονται, όταν καταλήξουμε στον σχηματισμό r/N ομάδων. Αυτό συμβαίνει και πάλι, γιατί θέλουμε να εντοπίσουμε την ομάδα που είναι όσο γίνεται πιο κοντά στο μοτίβο (από την άποψη της πληροφορίας που μας παρέχει για αυτό) και να δώσουμε τον πίνακα βάρους θέσης που την περιγράφει ως αρχικοποίηση στους αλγορίθμους μας.

Το μέτρο εγγύτητας (ή απόστασης) που χρησιμοποιούμε, τώρα, για τον υπολογισμό των αποστάσεων μεταξύ των ομάδων, βασίζεται στον υπολογισμό της λογαριθμικής πιθανοφάνειας της κάθε μίας που ορίζεται από τη σχέση

$$l(cl_j) = \ln P(X_i \in cl_j \mid \Theta) = \sum_{X_i \in cl_j} \left[\sum_{k=1}^{K} \sum_{r=1}^{M} \delta(X_{i_k}, a_r) \cdot \ln(\theta_{rk}^{cl_j}) \right]$$

$$= \sum_{k=1}^{K} \sum_{r=1}^{M} \eta_{rk}^{cl_j} \cdot \ln(\theta_{rk}^{cl_j}), \qquad (3.1)$$

όπου με $\eta_{rk}^{cl_j}$ συμβολίζουμε το πλήθος των εμφανίσεων του χαρακτήρα $a_r \in A$ στην $k-$στη θέση (για $k = 1,...,K$) σε όλες τις υπακολουθίες της ομάδας j, καθώς και της από κοινού λογαριθμικής πιθανοφάνειας όλων των ομάδων (ανά δύο μεταξύ τους) που ορίζεται ανάλογα με τη σχέση (3.1) και συγκεκριμένα, ως

$$l(cl_q \cup cl_j) = \ln P(X_i \in cl_q \cup cl_j \mid \Theta) = \sum_{X_i \in cl_q \cup cl_j} \left[\sum_{k=1}^{K} \sum_{r=1}^{M} \delta(X_{i_k}, a_r) \cdot \ln(\theta_{rk}^{cl_q \cup cl_j}) \right]$$

$$= \sum_{k=1}^{K} \sum_{r=1}^{M} \eta_{rk}^{cl_q \cup cl_j} \cdot \ln(\theta_{rk}^{cl_q \cup cl_j}) \,. \tag{3.2}$$

Έτσι, η απόσταση που ορίζουμε μεταξύ δύο ομάδων δίνεται, σύμφωνα με τις (4.1), (4.2), από τη σχέση

$$d(cl_q, cl_j) = l(cl_q) + l(cl_j) - l(cl_q \cup cl_j)\,,\ q \neq j, \tag{3.3}$$

και μάλιστα, ισχύει ότι $d(cl_q, cl_j) \geq 0$. Η τελευταία ποσότητα είναι μη αρνητική, διότι ο όρος $l(cl_q)$ εξαρτάται αποκλειστικά από τις υπακολουθίες της ομάδας q. Μετά τη συνένωση, όμως, των ομάδων q και j, όλες οι περιπτώσεις τους ανατίθενται στην καινούρια ομάδα, δηλαδή ισχύει ότι

$$\eta_{rk}^{cl_q \cup cl_j} = \eta_{rk}^{cl_q} + \eta_{rk}^{cl_j} \,. \tag{3.4}$$

Ακόμα, ο πίνακας $\theta^{cl_q \cup cl_j}$ που προκύπτει από τη συνένωση των ίδιων ομάδων και περιγράφει όλες τις υπακολουθίες που θα συμπεριλαμβάνει η νεοσυσταθείσα ομάδα, υπολογίζεται ως εξής

$$\theta^{cl_q \cup cl_j} = \frac{\theta^{cl_q} + \theta^{cl_j}}{\left|cl_q\right| + \left|cl_j\right|} \,. \tag{3.5}$$

Αξίζει να παρατηρήσουμε ότι στην περίπτωση της συνένωσης, σε κάποιο βήμα, δύο τυχαίων ομάδων q και j, οι αποστάσεις που ο αλγόριθμος απαιτεί να επαναϋπολογίσουμε αφορούν μόνο τις δύο αυτές ομάδες. Συγκεκριμένα, πρόκειται για τις αποστάσεις $d(cl_q, cl_t)$, για $t \neq q$ και $d(cl_j, cl_t)$, για $t \neq j$, όπου με τον συμβολισμό cl_t αναφερόμαστε σε όλες τις υπαρκτές ομάδες υπακολουθιών του τρέχοντος βήματος. Στο σημείο αυτό, κρίνεται αναγκαίο να δοθεί μεγάλη προσοχή, γιατί ο αλγόριθμος είναι ούτως ή άλλως υπολογιστικά δαπανηρός (και συνεπώς αργός). Επομένως, περιττοί υπολογισμοί θα τον επιβαρύνουν ακόμα περισσότερο.

Από τη στιγμή, τώρα, που σχηματίζονται οι r/N ομάδες (και ο συνθετικός αλγόριθμος τερματίζει), χρειαζόμαστε ένα κριτήριο επιλογής μιας ομάδας, η οποία θέλουμε να περιέχει όσο το δυνατόν μεγαλύτερη πληροφορία του μοτίβου. Υποθέτουμε, λοιπόν, ότι πρόκειται για την ομάδα εκείνη που οι υπακολουθίες της έχουν τα περισσότερα κοινά χαρακτηριστικά

μεταξύ τους (δηλαδή παρουσιάζουν τη μεγαλύτερη δυνατή συνοχή ή ομοιογένεια) σε σχέση με αυτά των υπολοίπων (οι οποίες ανήκουν στο background και είναι αναμενόμενο να μην παρουσιάζουν και μεγάλη ομοιότητα μεταξύ τους). Τότε, η λύση στο πρόβλημα αυτό παρέχεται από το μέτρο της *εντροπίας* που ελαχιστοποιείται για μια τέτοια ομάδα και στην περίπτωσή μας, ορίζεται για την ομάδα i ως

$$e_i = -\sum_{r=1}^{M}\sum_{k=1}^{K} \theta_{rk}^{cl_i} \log \theta_{rk}^{cl_i} \ , \ \ i = 1,...,r \,/\, N \,. \tag{3.6}$$

Ο πίνακας βάρους θέσης της τελευταίας ομάδας δίνεται ως αρχικοποίηση της παραμέτρου θ στους αλγορίθμους που χρησιμοποιούμε για την ανακάλυψη του μοτίβου. Όπως θα διαπιστώσουμε και πειραματικά στο 5ο κεφάλαιο, η αρχικοποίηση που παρέχεται από τον συνθετικό αλγόριθμο για την βασική παράμετρο του προβλήματος, οδηγεί σε μια πολύ καλύτερη προσέγγιση του μοτίβου από μια μέθοδο (και κυρίως από τον EM), σε σχέση με αυτήν που δίνει χρησιμοποιώντας τυχαία αρχικοποίηση για το ίδιο σύνολο δεδομένων.

Τα βήματα που ακολουθούνται για την εφαρμογή του συνθετικού αλγορίθμου περιγράφονται στον Πίνακα 3.4.

Πίνακας 3.4: Ο Συνθετικός Αλγόριθμος για το Πρόβλημα.

- Αρχικά, θεώρησε κάθε υπακολουθία X_i ως μια ομάδα, έστω cl_i, και υπολόγισε τον αντίστοιχο πίνακα βάρους θέσης θ^{cl_i} $(i = 1,2,...,r)$.
- Υπολόγισε τις ποσότητες $l(cl_i)$, $l(cl_i \cup cl_j)$ για κάθε $i, j \in \{1,...,r\}, i \neq j$, και με βάση αυτές, τις αποστάσεις $d(cl_i, cl_j)$ για κάθε $i, j \in \{1,...,r\}, i \neq j$.
- Συνένωσε τις ομάδες k και l, για τις οποίες ισχύει $d(cl_k, cl_l) = \min_{\substack{i,j=1,...,r \\ i \neq j}} [d(cl_i, cl_j)]$, σχηματίζοντας τη νέα ομάδα k.
- Επανάλαβε τα εξής βήματα:
  - Υπολόγισε τις ποσότητες $l(cl_k)$, $l(cl_k \cup cl_j)$ για κάθε υπαρκτή ομάδα $j \neq k$, και με βάση αυτές, τις νέες απαιτούμενες αποστάσεις $d(cl_k, cl_j)$ για κάθε $j \neq k$.
  - Συνένωσε τις ομάδες k και l, για τις οποίες ισχύει

$d(cl_k, cl_l) = \min_{\substack{i,j=1,\ldots,r \\ i \neq j}} [d(cl_i, cl_j)]$, σχηματίζοντας τη νέα ομάδα k

έως ότου το πλήθος των ομάδων γίνει ίσος με r / N.

ΚΕΦΑΛΑΙΟ 4. ΔΥΟ ΕΠΕΚΤΑΣΕΙΣ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΕΜ

4.1. Μειονεκτήματα του Αλγορίθμου ΕΜ

Μέχρι στιγμής, έχουμε υποθέσει ότι οι παρατηρήσεις του συνόλου εκπαίδευσης, στο πρόβλημα που μελετούμε, είναι ανεξάρτητες μεταξύ τους. Κάτι τέτοιο, όμως, στην πραγματικότητα δεν είναι αληθές. Πράγματι, εφόσον το σύνολο αυτό περιλαμβάνει όλες τις δυνατές υπακολουθίες μήκους K, υπάρχουν μέσα σε αυτό επικαλύψεις κάθε υπακολουθίας, οι οποίες διαφέρουν μεταξύ τους από ένα μέχρι $K-1$ σύμβολα. Για να γίνει αντιληπτός ο ισχυρισμός αυτός, μπορεί να ανατρέξει κανείς στο Σχήμα 4.1.

Σύμφωνα με τα παραπάνω και γνωρίζοντας ότι ο αλγόριθμος ΕΜ παρουσιάζει το μειονέκτημα της σύγκλισης σε κάποιο τοπικό μέγιστο της συνάρτησης Πιθανοφάνειας, είναι πολύ πιθανό αντί του πραγματικού μοτίβου, να ανακαλύψει ένα τμήμα του. Για παράδειγμα, υποθέτουμε ότι το βέλτιστο μοτίβο μήκους 5 μέσα σε ένα σύνολο συμβολοσειρών με χαρακτήρες από το αλφάβητο $A=\{A,B,E,G,M,Y,L\}$ είναι το $m=[MYBAG]$. Είναι πολύ πιθανό, λοιπόν, ο αλγόριθμος ΕΜ να δώσει ως μοτίβο την ακολουθία χαρακτήρων ‘ *LMYBA* ’. Όπως παρατηρούμε, σε αυτή την περίπτωση χάθηκε ένα πολύ σημαντικό ποσοστό (20%) της πληροφορίας που αναζητούμε. Ασφαλώς, θα μπορούσε να χαθεί ακόμα μεγαλύτερο ποσοστό (π.χ. 80% που αντιστοιχεί στην εύρεση ενός μόνο χαρακτήρα του

πραγματικού μοτίβου) και το αποτέλεσμα να μην είναι πια συμβατό με την πραγματικότητα (εφόσον οι υπόλοιποι 4 χαρακτήρες αντιστοιχούν σε θέσεις background).

Ένα ακόμα μειονέκτημα είναι η μεγάλη επίδραση της αρχικοποίησης. Όπως έχει παρατηρηθεί και μπορεί εύκολα να εξηγηθεί, η εκτιμούμενη λύση του μοτίβου από τον ΕΜ είναι σχεδόν αδύνατο να αποκλίνει από τη λύση που δέχεται αρχικά ως είσοδο ο αλγόριθμος (δηλαδή από τον πίνακα θ που παρέχεται ως αρχικοποίηση).

Το παραπάνω πρόβλημα είναι πιο έντονο στις περιπτώσεις, όπου το μοτίβο που ψάχνουμε περιέχει επαναλαμβανόμενους χαρακτήρες, π.χ. 'BABABA….', 'PAMEPAME….' κλπ. Τότε αυτό που συμβαίνει είναι ότι η μέθοδος, εξαιτίας της υπόθεσης της ανεξαρτησίας των παρατηρήσεων, εκτιμά ότι περισσότερες της μιας συμβολοσειρές (παρατηρήσεις) που περιέχουν τμήμα του πραγματικού επαναληπτικού μοτίβου είναι αντίγραφά του. Για παράδειγμα, έστω η συμβολοσειρά ….ABDCBABABABABACDCCB….. που περιλαμβάνει το πραγματικό μοτίβο 10 θέσεων 'BABABABABA'. Κατά την κατάτμηση της συμβολοσειράς (σε υπακολουθίες μήκους 10), παίρνουμε τις εξής υπο-ακολουθίες που αποτελούν και τις παρατηρήσεις του προβλήματος:

……

X_1= ABDCBABABA

X_2= BDCBABABAB

X_3= DCBABABABA

X_4= CBABABABAB

X_5= BABABABABA

X_6= ABABABABAC

X_7= BABABABACD

X_8= ABABABACDC

X_9= BABABACDCC

X_{10}= ABABACDCCB

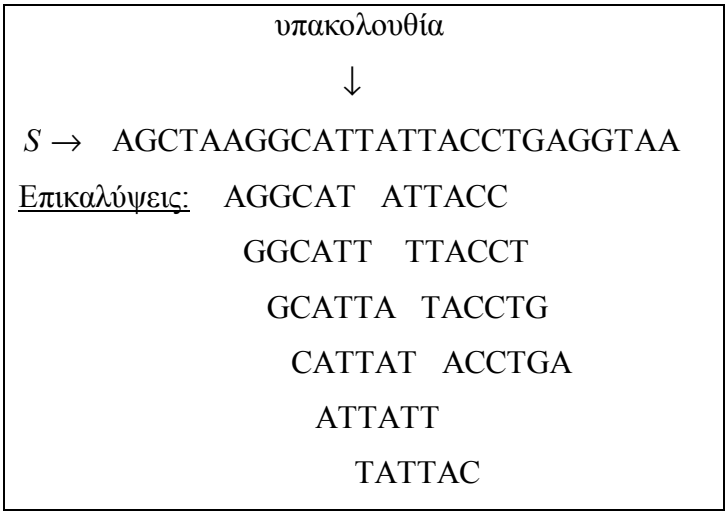
……

Όπως είναι φανερό, οι X_1, X_2, X_3, X_4, X_6, X_7, X_8, X_9, X_{10} περιέχουν μεν ένα σημαντικό ποσοστό της κρυμμένης πληροφορίας που αναζητούμε, αλλά η μέθοδος θα ήταν απόλυτα επιτυχής, εάν κατόρθωνε να ανιχνεύσει ως μοτίβο την υπακολουθία X_5. Αξίζει να σημειώσουμε ότι, μέσα στις προηγούμενες υπακολουθίες, το μοτίβο εμφανίζεται μετατοπισμένο κατά ένα πλήθος θέσεων είτε δεξιά είτε αριστερά.

Για τον λόγο αυτό, στο παρόν κεφάλαιο θα επιχειρήσουμε να εξαλείψουμε, όσο είναι δυνατόν, αυτή την αδυναμία του ΕΜ, εισάγοντας δύο παραλλαγές του. Οι μέθοδοι αυτές λαμβάνουν υπόψη τις επικαλύψεις των παρατηρήσεων και κατορθώνουν συχνά, όπως αποδεικνύεται πειραματικά στο $5^{ο}$ κεφάλαιο, να ανιχνεύσουν ένα μεγαλύτερο τμήμα του κρυμμένου μοτίβου και να υπερνικήσουν τον ΕΜ.

Πιο αναλυτικά, στην **πρώτη μέθοδο** χρησιμοποιούμε μια επέκταση του πίνακα θ που περιγράφει το μοτίβο και συγκεκριμένα προσθέτουμε σε αυτόν $K-1$ στήλες αριστερά και άλλες τόσες δεξιά του (Σχήμα 4.2). Εδώ, θα μπορούσε να θεωρηθεί ότι, κατά κάποιον τρόπο, ολισθαίνουμε κάθε υπακολουθία (παρατήρηση) που εξετάζουμε μέσα στον διευρυμένο πίνακα έτσι, ώστε να ανακαλύψουμε μια περιοχή με συνεχόμενες θέσεις – στήλες του πίνακα (αν φυσικά υπάρχει) που να ταιριάζει βέλτιστα σε αυτό. Αυτό που προσπαθούμε να επιτύχουμε, με τη συγκεκριμένη μέθοδο, είναι να μην υποπέσει ο αλγόριθμος ΕΜ στο «σφάλμα» του εντοπισμού μόνο ενός τμήματος του μοτίβου (π.χ. 3 θέσεις από το μοτίβο και 3 θέσεις που αντιστοιχούν στο background), αλλά να κατορθώσει να εντοπίσει όλα τα σύμβολα που το περιγράφουν.

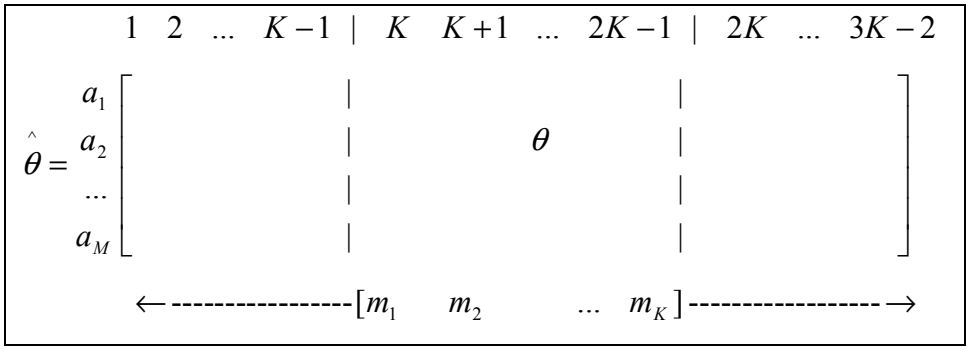
Η **δεύτερη μέθοδος** είναι στην ουσία μία παραλλαγή της πρώτης. Εδώ, ο πίνακας θ είναι σταθερός (K θέσεων-στηλών), αλλά ενδέχεται μόνο ένα τμήμα του υπό εξέταση δεδομένου (υπο-ακολουθίας) να καλύπτεται από αυτόν. Θα μπορούσαμε να ισχυριστούμε ότι μοιάζει σαν να ολισθαίνουμε τον πίνακα θ πάνω στην υποακολουθία προκειμένου να βρούμε το μέγιστο δυνατό (μεταξύ τους) ταίριασμα (το μέγιστο πλήθος στηλών του και τις θέσεις του) είτε από δεξιά είτε από αριστερά, το οποίο αντιστοιχεί στο μέγιστο δυνατό τμήμα του μοτίβου. Για να είναι, βέβαια, κάτι τέτοιο εφικτό δεν αρκούν μόνο δύο είδη κατανομών (μία για το μοτίβο και μία για το background), όπως θα διαπιστώσουμε στις ενότητες που ακολουθούν.



Σχήμα 4.1: Όλες οι δυνατές επικαλύψεις μιας υπακολουθίας μήκους 6 με χαρακτήρες από το αλφάβητο A={A,G,C,T}.

4.2. Ο Αλγόριθμος του Εκτεταμένου Πίνακα

Με τις γνωστές υποθέσεις, θεωρούμε τώρα ότι το μοτίβο δεν περιγράφεται από έναν MxK διάστασης πίνακα βάρους θέσης $\theta = [\theta_{jk}]$, αλλά από έναν πίνακα $\hat{\theta} = [\hat{\theta}_{jk}]$ διάστασης $Mx(3K-2)$. Στον τελευταίο πίνακα, υποθέτουμε ότι το τμήμα του εκείνο που αντιστοιχεί στον πίνακα του αλγορίθμου ΕΜ βρίσκεται στο κέντρο του, δηλαδή παριστάνεται από τις στήλες $K, K+1,...,2K-1$. Βέβαια, στη συνέχεια και μετά τη σύγκλιση του αλγορίθμου εκπαίδευσης, είναι δυνατό να βρεθεί ότι οι K στήλες του πίνακα που περιγράφουν το μοτίβο δεν βρίσκονται στο κέντρο αλλά είναι μετατοπισμένες προς τα αριστερά ή δεξιά. Για πληρέστερη κατανόηση των παραπάνω, μπορεί κανείς να παρατηρήσει το Σχήμα 4.2.



Σχήμα 4.2: Η ολίσθηση του μοτίβου $[m_1 m_2 ... m_K]$ στον πίνακα $\hat{\theta}$ διάστασης $Mx(3K-2)$.

Ο λόγος, για τον οποίο χρησιμοποιούμε έναν πίνακα με σχεδόν τριπλάσιο πλήθος στηλών, οφείλεται στην προσπάθειά μας να εντοπίσουμε όσο το δυνατόν περισσότερες θέσεις του πραγματικού μοτίβου. Συγκεκριμένα, στην περίπτωση που από τον πίνακα θ υπολείπονται κάποιες θέσεις του μοτίβου είτε αριστερά είτε δεξιά, υποθέτουμε ότι αυτές έχουν μετακινηθεί, αντίστοιχα, είτε μέσα στις πρώτες $K-1$ στήλες του καινούριου πίνακα $\hat{\theta}$ είτε στις τελευταίες $K-1$ στήλες του. Πρόκειται, λοιπόν, για μια ολίσθηση του μοτίβου μέσα στον $\hat{\theta}$ έτσι, ώστε να εντοπίσει τη διαδοχική $K-$άδα στηλών που του αντιστοιχεί.

Μια σημαντική παρατήρηση, που πρέπει να τονιστεί ιδιαίτερα, είναι ότι μέσω του εκτεταμένου πίνακα πετυχαίνουμε έμμεσα την εισαγωγή της χωρικής πληροφορίας στο μοντέλο περιγραφής του μοτίβου, καθώς επιτρέπουμε την ολίσθηση κάθε παρατηρούμενης υποακολουθίας μέσα στον διευρυμένο πίνακα με σκοπό να πετύχουμε βέλτιστο ταίριασμα. Στην ουσία, είναι σαν να έχουμε (έμμεσα) εφαρμόσει χωρικούς τελεστές για να ανιχνεύσουμε όσο το δυνατόν μεγαλύτερο τμήμα του μοτίβου και στην ιδανική περίπτωση, όλο το μοτίβο.

Τα πειράματα που πραγματοποιήσαμε απέδειξαν ότι δεν αποκλείεται μέσα στον συνήθη πίνακα θ του ΕΜ να έχει εμφανιστεί ακόμα και μία μόνο θέση του πραγματικού μοτίβου και συγκεκριμένα, είτε η τελευταία του θέση στην 1^{η} στήλη του θ είτε η πρώτη του θέση στην τελευταία στήλη του ίδιου πίνακα. Τότε, στη μεν πρώτη περίπτωση τα υπόλοιπα γράμματα του μοτίβου ενδέχεται να βρεθούν μέσα στις πρώτες $K-1$ στήλες του $\hat{\theta}$, στη δε δεύτερη μέσα στις $K-1$ τελευταίες στήλες του ίδιου πίνακα.

Για παράδειγμα, έστω ότι ο κλασικός αλγόριθμος ΕΜ επιστρέφει, μέσω του τελικού πίνακα θ το μοτίβο '*RAHME*' (δηλαδή, εδώ, είναι $K = 5$), ενώ το βέλτιστο είναι '*HMERA*'. Τότε, λαμβάνοντας το μέγιστο στοιχείο της $10^{ης}$ και $11^{ης}$ στήλης του $\hat{\theta}$, αναμένουμε αυτές να αντιστοιχούν στα γράμματα R και A αντίστοιχα, δηλαδή σε αυτά που δεν βρήκε ο ΕΜ (Σχήμα 4.3), καθώς εδώ έχει πραγματοποιηθεί μια ολίσθηση του μοτίβου κατά 2 θέσεις δεξιά μέσα στον πίνακα $\hat{\theta}$. Άρα, αυτό εντοπίζεται μέσα στον πίνακα μεταξύ των στηλών 7 (= (K −1) + 2) και 11 (= K + (K −1) + 2).

$\hat{\theta} =$

	1	2	3	4	5	6	7	8	9	10	11	12	13
A	.03	.3	.03	.1	.1	.5	.2	.1	.11	0	.8	.3	.1
E	.07	.2	.18	.2	.13	0	0	.1	.64	0	.0	.1	.1
H	.22	2	.12	.4	0	.1	.8	0	.07	0	0	.1	.2
J	.13	.1	.38	0	.17	0	0	0	.06	0	.1	.5	.2
M	.15	.1	.12	.2	0	.4	0	.7	.03	.1	0	0	.4
R	.4	.1	.17	.1	.6	0	0	.1	.09	.9	.1	0	0
	R	*A*	*J*	*H*	*R*	*A*	*H*	*M*	*E*	*R*	*A*	*J*	*M*

↑ θέση έναρξης ↑ θέση λήξης μοτίβου

Σχήμα 4.3: Ο πίνακας αναπαράστασης του μοτίβου '*HMERA*' στην $1^{η}$ επέκταση του αλγορίθμου ΕΜ.

Σύμφωνα με όσα αναφέρθηκαν προηγουμένως, μπορεί να φανταστεί κανείς ότι ο πίνακας $\hat{\theta}$ περιέχει $2K-1$ δυνατές θέσεις έναρξης του μοτίβου, αφού αυτό μπορεί να ξεκινήσει από την $1^{η}$ στήλη του (και να εμφανίζεται στις πρώτες K στήλες του) μέχρι και την $(2K-1)^{η}$ (και να εμφανίζεται από την $(2K-1)^{η}$ έως την $(3K-2)^{η}$). Ανάλογα, βέβαια, με το ποιες είναι οι K διαδοχικές στήλες που αποτελούν τις θέσεις μοτίβου, οι υπόλοιπες $2K-2$ στήλες του αντιπροσωπεύουν θέσεις background μέσα στις συμβολοσειρές.

Συμβολίζουμε, τώρα, με $\hat{\theta}_j$ τον MxK πίνακα που περιλαμβάνει τις στήλες του $\hat{\theta}$ από την $j-$στη μέχρι την $(j+K-1)-$στη, όπως αυτός ορίστηκε στον αλγόριθμο ΕΜ. Τότε, στον πίνακα 4.1, παρουσιάζεται αναλυτικά για κάθε στήλη του $\hat{\theta}$ σε πόσους και σε ποιους συγκεκριμένα από τους MxK πίνακες $\left\{\hat{\theta}_j\right\}_{j=1}^{2K-1}$ ενδέχεται να εμφανιστεί (ως μία από τις θέσεις του μοτίβου).

Πίνακας 4.1: Συμμετοχή κάθε Στήλης του $Mx(3K-2)$ Πίνακα $\hat{\theta}$ για την Αναπαράσταση του Μοτίβου.

Στήλη $k \in \{1,..,3K-2\}$ του $\hat{\theta}$	Πιθανοί πίνακες εμφάνισης $\hat{\theta}_j$, $j \in \{1,2,...,2K-1\}$	Συνολικός αριθμός δυνατών εμφανίσεων
1	$\hat{\theta}_j,\ j=1$	1
2	$\hat{\theta}_j,\ j=1,2$	2
…	…	…
$K-1$	$\hat{\theta}_j,\ j=1,2,...,K-1$	$K-1$
K	$\hat{\theta}_j,\ j=1,2,...,K$	K
$K+1$	$\hat{\theta}_j,\ j=2,3,...,K+1$	K
…	…	…
$2K-1$	$\hat{\theta}_j,\ j=K,K+1,...,2K-1$	K
$2K$	$\hat{\theta}_j,\ j=K+1,...,2K-1$	$K-1$
$2K+1$	$\hat{\theta}_j,\ j=K+2,...,2K-1$	$K-2$
…	…	…
$3K-2$	$\hat{\theta}_j,\ j=2K-1$	1

Πίνακας 4.2: Συνοπτική Περιγραφή του Πίνακα 4.1.

Στήλη $k \in \{1,..,3K-2\}$ του $\hat{\theta}$	Πιθανοί πίνακες εμφάνισης $\left\{\hat{\theta}_j\right\}_{j=1}^{2K-1}$	Συνολικός αριθμός δυνατών εμφανίσεων
$k = 1:K-1$	$j = 1:k$	k
$k = K:2K-1$	$j = k-K+1:k$	K
$k = 2K:3K-2$	$j = k-K+1:2K-1$	$3K-k-1$

Οι παραπάνω Πίνακες δηλώνουν ότι ο πίνακας $\hat{\theta}$ «κρύβει» $2K-1$ διαφορετικές κατανομές που ακολουθούν οι υπακολουθίες X_i που είναι μοτίβο, καθώς και εκείνες που περιλαμβάνουν ένα τμήμα του (από μία μέχρι $K-1$ θέσεις του). Έτσι, υποθέτουμε την ύπαρξη μίας «κρυμμένης» μεταβλητής Z_i για κάθε υπακολουθία X_i τέτοια, ώστε να ισχύει $Z_i = j$, εάν η X_i ακολουθεί την κατανομή $\hat{\theta}_j$, $j \in \{1,2,...,2K-1\}$ (του μοτίβου), και $Z_i = 2K$, αν η X_i ανήκει στο background. Με άλλα λόγια, η κρυμμένη μεταβλητή προσδιορίζει το μέγεθος της ολίσθησης του πίνακα πάνω στην υποακολουθία, ενώ αν η τιμή της είναι $2K$, τότε δεν υπάρχει ολίσθηση. Έτσι, για το σύνολο των παρατηρήσεων $\{X_i\}_{i=1}^{r}$, η λύση του προβλήματος που εξετάζουμε δίδεται από τα ζεύγη $\{X_i, Z_i\}_{i=1}^{r}$. Μη γνωρίζοντας, όμως, την τιμή των μεταβλητών Z_i και αφού το εν λόγω μοντέλο αποτελείται από $2K$ το πλήθος κατανομές, υποθέτουμε ότι ένα δείγμα X_i ακολουθεί την κατανομή του μοτίβου $\hat{\theta}_j$ με μία εκ των προτέρων πιθανότητα $\pi_j = P(Z_i = j)$, $j \in \{1,2,...,2K-1\}$, και την κατανομή του background με μία εκ των προτέρων πιθανότητα $\pi_{2K} = 1 - \sum_{j=1}^{2K-1} \pi_j = P(Z_i = 2K)$.

Συνεπώς, τώρα, το σύνολο των παραμέτρων του μοντέλου είναι το $\Psi = \{\hat{\theta}, \vec{\pi}, b\}$, όπου $\vec{\pi} = [\pi_1, \pi_2, ..., \pi_{2K}]$ και b το γνωστό διάνυσμα (από τον ΕΜ) που αφορά την κατανομή background για καθένα από τα σύμβολα του αλφάβητου. Ακόμα, συμβολίζουμε με

$\Theta = \{\hat{\theta}, b\}$. Προφανώς, σε αυτή την περίπτωση, η συνάρτηση πιθανότητας για κάθε δείγμα X_i $(i = 1,2,...,r)$, υπολογίζεται από το μικτό μοντέλο

$$f(X_i \mid \Psi) = \sum_{j=1}^{2K-1} \pi_j \cdot \rho(X_i \mid \hat{\theta}) + (1 - \sum_{j=1}^{2K-1} \pi_j) \cdot \rho(X_i \mid b). \tag{4.1}$$

Και πάλι, υποθέτουμε ότι οι θέσεις σε κάθε υπακολουθία είναι ανεξάρτητες μεταξύ τους, οπότε οι συναρτήσεις κατανομής για την κατανομή του *μοτίβου* και του *background* είναι αντίστοιχα

$$\rho(X_i \mid \hat{\theta}) = P(X_i \mid Z_i = j, \Theta) = \prod_{k=1}^{K} \hat{\theta}_{X_{i_k}} = \prod_{k=1}^{K} \prod_{l=1}^{M} \theta_{l,j+k-1}^{\delta(X_{i_k}, a_l)}, \; j = 1,...,2K-1 \tag{4.2}$$

και

$$\rho(X_i \mid b) = P(X_i \mid Z_i = 2K, \Theta) = \prod_{k=1}^{K} b_l = \prod_{k=1}^{K} \prod_{l=1}^{M} b_l^{\delta(X_{i_k}, a_l)} = \prod_{l=1}^{M} b_l^{\sum_{k=1}^{K} \delta(X_{i_k}, a_l)}, \tag{4.3}$$

όπου με δ συμβολίζουμε τη δείκτρια συνάρτηση $\delta(X_{i_k}, a_l) = \begin{cases} 1, X_{i_k} = a_l \\ 0, X_{i_k} \neq a_l \end{cases}$.

Σε αυτή την περίπτωση, θέλουμε να μεγιστοποιήσουμε τη λογαριθμική συνάρτηση (πιθανοφάνειας)

$$L(X) = \ln P(X \mid \Psi) = \ln[P(X_1 \mid \Psi) \cdot P(X_2 \mid \Psi)...P(X_r \mid \Psi)] = \sum_{i=1}^{r} \ln f(X_i \mid \Psi)$$

$$= \sum_{i=1}^{r} \ln \left[\sum_{j=1}^{2K-1} \pi_j \cdot \rho(X_i \mid \hat{\theta}) + \left(1 - \sum_{j=1}^{2K-1} \pi_j \right) \cdot \rho(X_i \mid b) \right]. \tag{4.4}$$

Αν συμβολίσουμε, τώρα, με $\Phi_j(X_i \mid \Theta) = P(X_i \mid Z_i = j, \Theta)$, $j \in \{1,2,...,2K\}$, τότε όπως και στο Expectation – βήμα του ΕΜ, ο αλγόριθμος υπολογίζει τις εκ των υστέρων πιθανότητες

$$P(Z_i = j \mid X_i, \Psi) = \frac{\pi_j \cdot \Phi_j(X_i \mid \Theta)}{\sum_{l=1}^{2K} \pi_l \cdot \Phi_l(X_i \mid \Theta)}, j = 1,...,2K, \tag{4.5}$$

καθώς, εδώ, υπάρχουν $2K$ κατανομές που μπορούν να ακολουθούν οι παρατηρήσεις μας.

Έτσι, η μέση (πλήρης) λογαριθμική πιθανοφάνεια, δίδεται από τη σχέση

$$Q = \sum_{i=1}^{r} \left[\ln P(X_i, Z_i \mid \Psi)\right] \cdot P(Z_i \mid X_i, \Psi) = \sum_{i=1}^{r}\sum_{j=1}^{2K} P(Z_i = j \mid X_i, \Psi) \cdot \ln P(X_i, Z_i = j \mid \Psi)$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{2K} P(Z_i = j \mid X_i, \Psi) \cdot \ln\left[P(Z_i = j) \cdot P(X_i \mid Z_i = j, \Psi)\right]$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{2K} P(Z_i = j \mid X_i, \Psi) \cdot \ln\left[\pi_j \cdot \Phi_j(X_i \mid \Theta)\right] \Rightarrow$$

$$Q = \sum_{i=1}^{r}\sum_{j=1}^{2K} P(Z_i = j \mid X_i, \Psi) \cdot \ln \pi_j + \sum_{i=1}^{r}\sum_{j=1}^{2K} P(Z_i = j \mid X_i, \Psi) \cdot \ln \Phi_j(X_i \mid \Theta). \quad (4.6)$$

Συνεχίζοντας όπως και στο Maximization-βήμα του EM, ο αλγόριθμος υπολογίζει με βάση την παραπάνω σχέση για την Q, τις νέες παραμέτρους $\Psi = \left\{(\pi_j)_{j=1}^{2K}, \hat{\theta}, b\right\}$, τις οποίες και χρησιμοποιεί στην επόμενη επανάληψη. Συγκεκριμένα, έχουμε ένα πρόβλημα μεγιστοποίησης με περιορισμούς που προκύπτουν εξαιτίας της στοχαστικής φύσης των μεταβλητών του προβλήματος. Στην προκειμένη περίπτωση, οι νέες εκ των προτέρων πιθανότητες υπολογίζονται από τους τύπους

$$\pi_j = \frac{1}{r}\left(\sum_{i=1}^{r} P(Z_i = j \mid X_i, \Psi)\right),\ j = 1,...,2K, \quad (4.7)$$

οι οποίοι προκύπτουν χρησιμοποιώντας τη σχέση

$$Q_\pi = \sum_{i=1}^{r}\sum_{j=1}^{2K} P(Z_i = j \mid X_i, \Psi) \cdot \ln \pi_j$$

και τον περιορισμό $\sum_{j=1}^{2K} \pi_j = 1$.

Χρησιμοποιώντας, στη συνέχεια, τον δεύτερο όρο (έστω Q_Φ) του αθροίσματος της σχέσης (4.6), θέτουμε $Q_\Phi = Q_{\hat{\theta}} + Q_b$, όπου

$$Q_{\hat{\theta}} = \sum_{i=1}^{r}\sum_{j=1}^{2K-1} P(Z_i = j \mid X_i, \Psi) \cdot \ln\left[\prod_{k=1}^{K}\prod_{l=1}^{M} \theta_{l,j+k-1}^{\ \delta(X_{ik}, a_l)}\right]$$

$$= \sum_{i=1}^{r} \sum_{j=1}^{2K-1} P(Z_i = j \mid X_i, \Psi) \cdot \sum_{k=1}^{K} \sum_{l=1}^{M} \delta(X_{i_k}, a_l) \cdot \ln(\theta_{l,j+k-1}) \tag{4.8}$$

και

$$Q_b = \sum_{i=1}^{r} P(Z_i = 2K \mid X_i, \Psi) \cdot \ln\left[\prod_{k=1}^{K} \prod_{l=1}^{M} b_l^{\delta(X_{i_k}, a_l)} \right]$$

$$= \sum_{i=1}^{r} P(Z_i = 2K \mid X_i, \Psi) \cdot \sum_{k=1}^{K} \sum_{l=1}^{M} \delta(X_{i_k}, a_l) \cdot \ln(b_l) . \tag{4.9}$$

Για να υπολογίσουμε τον πίνακα $\hat{\theta} = [\hat{\theta}_{lm}] \in \Re^{Mx(3K-2)}$ του επόμενου βήματος, λαμβάνουμε υπόψη τους περιορισμούς $\sum_{l=1}^{M} \hat{\theta}_{lm} = 1$, για $m = 1,2,...,3K-2$. Επιπλέον, θέτουμε $m = j + k - 1,$ οπότε $k = m - j + 1$. Θεωρούμε, τώρα, τους πολλαπλασιαστές Lagrange λ_m, $m = 1,2,...,K$. Παραγωγίζοντας ως προς $\hat{\theta}_{lm}$ και θέτοντας ίση με μηδέν τη συνάρτηση

$$Q_1 = \sum_{i=1}^{r} \sum_{j=1}^{m} P(Z_i = j \mid X_i, \Psi) \cdot \sum_{m=1}^{K} \sum_{l=1}^{M} \delta(X_{i_{m-j+1}}, a_l) \cdot \ln(\hat{\theta}_{lm}) + \sum_{m=1}^{K} \lambda_m \cdot \left(1 - \hat{\theta}_{lm}\right), \tag{4.10}$$

έπεται ότι

$$\hat{\theta}_{lm} = \frac{\sum_{i=1}^{r} \sum_{j=1}^{m} P(Z_i = j \mid X_i, \Psi) \cdot \delta(X_{i_{m-j+1}}, a_l)}{\sum_{i=1}^{r} \sum_{j=1}^{m} P(Z_i = j \mid X_i, \Psi)}, l = 1,...,M,\ m \in \{1,...,K-1\}. \tag{4.11}$$

Παρόμοια, προκύπτει ότι

$$\hat{\theta}_{lm} = \frac{\sum_{i=1}^{r} \sum_{j=m-K+1}^{m} P(Z_i = j \mid X_i, \Psi) \cdot \delta(X_{i_{m-j+1}}, a_l)}{\sum_{i=1}^{r} \sum_{j=m-K+1}^{m} P(Z_i = j \mid X_i, \Psi)}, l = 1,...,M,\ m \in \{K,...,2K-1\} \tag{4.12}$$

και

$$\hat{\theta}_{lm} = \frac{\sum_{i=1}^{r}\sum_{j=m-K+1}^{2K-1} P(Z_i = j \mid X_i, \Psi)\cdot \delta(X_{i_{m-j+1}}, a_l)}{\sum_{i=1}^{r}\sum_{j=m-K+1}^{2K-1} P(Z_i = j \mid X_i, \Psi)}, l = 1,...,M,\ m \in \{2K,...,3K-2\}. \quad (4.13)$$

Για τον υπολογισμό, τώρα, του νέου διανύσματος b, θεωρούμε τον πολλαπλασιαστή Lagrange λ (λόγω του περιορισμού $\sum_{l=1}^{M} b_l = 1$). Παραγωγίζοντας, λοιπόν, ως προς b_l (και θέτοντας ίση με μηδέν) τη συνάρτηση

$$Q_2 = \sum_{i=1}^{r} P(Z_i = 2K \mid X_i, \Psi)\cdot \sum_{k=1}^{K}\sum_{l=1}^{M} \delta(X_{i_k}, a_l)\cdot \ln(b_l) - \lambda \cdot \left(1 - \sum_{l=1}^{M} b_l\right), \quad (4.14)$$

και λαμβάνοντας υπόψη ότι $\sum_{j=1}^{M}\sum_{k=1}^{K} \delta(X_{i_k}, a_j) = K$, το νέο διάνυσμα b υπολογίζεται από τις σχέσεις

$$b_l = \frac{\sum_{i=1}^{r} P(Z_i = 2K \mid X_i, \Psi)\cdot \sum_{k=1}^{K} \delta(X_{i_k}, a_l)}{K \cdot \sum_{i=1}^{r} P(Z_i = 2K \mid X_i, \Psi)}, l = 1,...,M. \quad (4.15)$$

Συνοψίζοντας τα παραπάνω αποτελέσματα, η μέθοδος του Εκτεταμένου πίνακα (Extended array) για το πρόβλημα που εξετάζουμε, είναι η ακόλουθη:

Πίνακας 4.3: Ο Αλγόριθμος του Εκτεταμένου πίνακα (Extended Array).

- Αρχικά: Θεώρησε τυχαίες τιμές για τις παραμέτρους $\Psi^{(\pi)}$.
- Επανάλαβε τα παρακάτω
  - <u>*Expectation – βήμα:*</u> Υπολόγισε τις εκ των υστέρων πιθανότητες

$$P(Z_i = j \mid X_i, \Psi) = \frac{\pi_j \cdot \Phi_j(X_i \mid \Theta)}{\sum_{l=1}^{2K} \pi_l \cdot \Phi_l(X_i \mid \Theta)},\ i = 1,...,r,\ j = 1,2,...,2K.$$

- *Maximization – βήμα:* Εκτίμησε τις νέες παραμέτρους $\Psi^{(\nu)} = \arg\max_{\Psi} Q(\Psi, \Psi^{(\pi)})$,

όπου

$$Q(\Psi, \Psi^{(\pi)}) = \sum_{i=1}^{r} \sum_{j=1}^{2K} P(Z_i = j \mid X_i, \Psi) \cdot \ln P(X_i, Z_i = j \mid \Psi),$$

από τις σχέσεις:

1) $\pi_j = \frac{1}{r}\left(\sum_{i=1}^{r} P(Z_i = j \mid X_i, \Psi) \right),\ j = 1,2,...,2K$

2) $\hat{\theta}_{lm} = \begin{cases} \dfrac{\sum_{i=1}^{r} \sum_{j=1}^{m} P(Z_i = j \mid X_i, \Psi) \cdot \delta(X_{i_{m-j+1}}, a_l)}{\sum_{i=1}^{r} \sum_{j=1}^{m} P(Z_i = j \mid X_i, \Psi)}, & l = 1,...,M,\ m \in \{1,...,K-1\} \\ \dfrac{\sum_{i=1}^{r} \sum_{j=m-K+1}^{m} P(Z_i = j \mid X_i, \Psi) \cdot \delta(X_{i_{m-j+1}}, a_l)}{\sum_{i=1}^{r} \sum_{j=m-K+1}^{m} P(Z_i = j \mid X_i, \Psi)}, & l = 1,...,M,\ m \in \{K,...,2K-1\} \\ \dfrac{\sum_{i=1}^{r} \sum_{j=m-K+1}^{2K-1} P(Z_i = j \mid X_i, \Psi) \cdot \delta(X_{i_{m-j+1}}, a_l)}{\sum_{i=1}^{r} \sum_{j=m-K+1}^{2K-1} P(Z_i = j \mid X_i, \Psi)}, & l = 1,...,M,\ m \in \{2K,...,3K-2\} \end{cases}$

3) $b_l = \dfrac{\sum_{i=1}^{r} P(Z_i = 2K \mid X_i, \Psi) \cdot \sum_{k=1}^{K} \delta(X_{i_k}, a_l)}{K \cdot \sum_{i=1}^{r} P(Z_i = 2K \mid X_i, \Psi)}, l = 1,...,M,$

όσο ισχύει $Q(\Psi, \Psi^{(\nu)}) > Q(\Psi, \Psi^{(\pi)}) + \varepsilon$, όπου $\varepsilon > 0$.

Συγκεκριμένα, αφού ληφθούν, μετά από κάποιο πλήθος επαναλήψεων των δύο βασικών βημάτων του αλγορίθμου, οι τελικές παράμετροι, αναζητούμε μέσα στον πίνακα $\hat{\theta}$ το

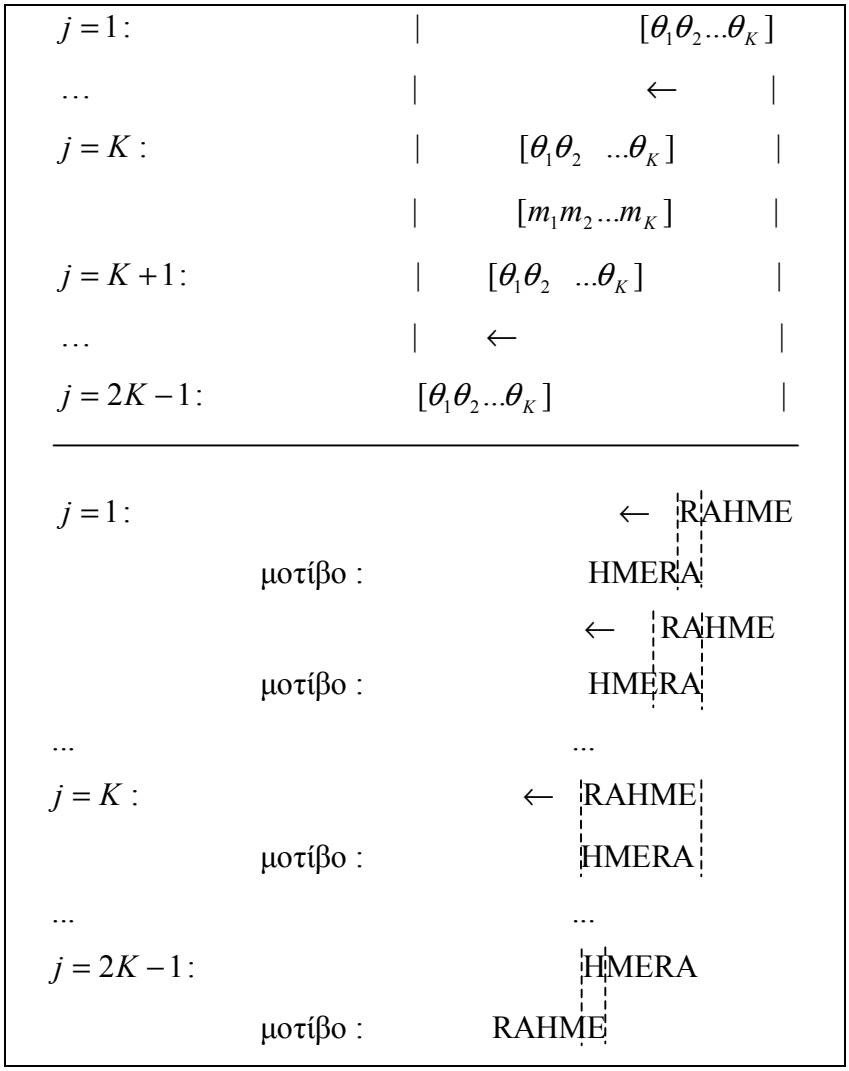
μέγιστο στοιχείο κάθε στήλης του $(\max_i)_{i=1}^{3K-2}$. Κατόπιν, ορίζουμε τις ποσότητες $(s_j)_{j=1}^{2K-1}$, όπου $s_j = \sum_{i=j}^{j+K-1} \max_i$, και βρίσκοντας τη μέγιστη $s_t = \max_{j=1,...,2K-1}(s_j)$ από αυτές, θεωρούμε ότι το μοτίβο αντιστοιχεί στην $t-$στη από τις $2K-1$ δυνατές κατανομές (δηλαδή στη Φ_t). Συνεπώς, τα σύμβολα από τα οποία σχηματίζεται το μοτίβο εξάγονται, όπως και στον ΕΜ, από τη γραμμή του MxK πίνακα $\hat{\theta}_t$ στην οποία αντιστοιχεί το μέγιστο στοιχείο κάθε στήλης του. Τέλος, οι υπακολουθίες που θεωρούνται ως αντίγραφα του μοτίβου λαμβάνονται από τις μέγιστες εκ των υστέρων πιθανότητες του Expectation – βήματος, οι οποίες αντιστοιχούν στην $t-$στη κατανομή.

4.3. Ο Αλγόριθμος του Οριοθετούμενου Πίνακα

Σε αυτή τη δεύτερη προσέγγιση, ο πίνακας αναπαράστασης θ του μοτίβου εξακολουθεί (όπως στον ΕΜ) να έχει διάσταση MxK. Όπως και στην πρώτη παραλλαγή του, υποθέτουμε με το ίδιο σκεπτικό ότι υπάρχουν $2K-1$ δυνατές κατανομές, τις οποίες ενδέχεται να ακολουθούν οι υπακολουθίες που αποτελούν το μοτίβο. Παρ' όλ' αυτά, η προσέγγιση αυτή υποθέτει ότι μόνο ένα τμήμα του πραγματικού μοτίβου μπορεί να περιγράφεται από τον πίνακα πιθανοτήτων. Το ερώτημα, βέβαια, που προκύπτει είναι πώς θα επιλέξουμε αυτό το τμήμα του θ και σε ποιο κομμάτι του μοτίβου ταιριάζει καλύτερα (ισοδύναμα ποια από τις $2K-1$ κατανομές αντιστοιχεί στο μοτίβο). Για να απαντηθεί κάτι τέτοιο, επιβάλλεται πρώτα να δώσουμε κάποιες διευκρινίσεις για τις κατανομές αυτές.

Εδώ, θεωρούμε ότι η μεταβλητή j ($j \in \{1,2,...,K\}$) αντιστοιχεί στο πλήθος των τελευταίων θέσεων μιας υπακολουθίας $X_i = (X_{i_1}, X_{i_2}, ..., X_{i_K})$, οι οποίες αποτελούν θέσεις μοτίβου (j πρώτες στήλες του θ), ενώ οι πρώτες $K-j$ θέσεις του ανήκουν στο background (το ίδιο ισχύει και για τις $K-j$ τελευταίες στήλες του θ). Παρόμοια, για $j \in \{K+1,...,2K-1\}$, ισχυριζόμαστε ότι οι $2K-j$ πρώτες θέσεις της X_i είναι θέσεις μοτίβου (και εμφανίζονται στις $2K-j$ τελευταίες στήλες του θ) και οι υπόλοιπες $j-K$ αντιστοιχούν σε θέσεις background (πρώτες $j-K$ στήλες του θ). Η μέθοδος αυτή και οι τύποι που θα προκύψουν για το σύνολο των παραμέτρων της αναπαριστούν μία νοητή

ολίσθηση του MxK πίνακα $\theta = [\theta_1 \theta_2 ... \theta_K]$ (όπου θ_i η $i-$στη στήλη του θ), από δεξιά προς τα αριστερά, πάνω στο μοτίβο, όπως φαίνεται στο σχήμα 4.4. Επειδή η σύλληψη της παραπάνω υπόθεσης είναι φυσικό να μην καταστεί σαφής με βάση την προηγούμενη περιγραφή, ακολουθεί και ένα συγκεκριμένο παράδειγμα, μετά την γενική περιγραφή που παρέχεται μέσω του σχήματος.



Σχήμα 4.4: Η ολίσθηση του MxK πίνακα $\theta = [\theta_1 \theta_2 ... \theta_K]$ πάνω στο μοτίβο $[m_1 m_2 ... m_K]$.

Όπως και στην προηγούμενη μέθοδο, για $j = 2K$, η X_i ανήκει στο background. Έτσι, για το σύνολο των παραμέτρων $\Psi = \left\{ \vec{\pi}, \theta, b \right\}$ με $\Theta = \{\theta, b\}$, (το $\pi = (\pi_1, ..., \pi_{2K})$ εκφράζει ό,τι και στην προηγούμενη μέθοδο) ορίζουμε τις συναρτήσεις πιθανότητας

$$\rho(X_i \mid j,\Theta)=\begin{cases}\left[\prod_{l=1}^{M} b_l^{\sum_{k=1}^{K-j}\delta(X_{i_k},a_l)}\right]\cdot\left[\prod_{k=1}^{j}\prod_{l=1}^{M}\theta_{lk}^{\delta(X_{i_{k+K-j}},a_l)}\right], & j\in\{1,...,K\}\\ \left[\prod_{l=1}^{M}\prod_{k=1}^{2K-j}\theta_{l,j-K+k}^{\delta(X_{i_k},a_l)}\right]\cdot\left[\prod_{l=1}^{M} b_l^{\sum_{k=2K-j+1}^{K}\delta(X_{i_k},a_l)}\right], & j\in\{K+1,...,2K-1\}\\ \prod_{l=1}^{M} b_l^{\sum_{k=1}^{K}\delta(X_{i_k},a_l)}, & j=2K\end{cases} \qquad (4.16).$$

Και εδώ, υπάρχει μία κρυμμένη μεταβλητή Z_i για κάθε X_i που φανερώνει την κατανομή που ακολουθεί ή αλλιώς περιγράφει το πλήθος των θέσεων ολίσθησης του πίνακα πάνω στην υπό εξέταση υπακολουθία (για $Z_i = 2K$ σημαίνει ότι δεν θα υπάρξει ολίσθηση). Στον πίνακα 4.4, μπορεί να διακρίνει κανείς σε ποιες και σε πόσες από τις διαφορετικές κατανομές είναι δυνατόν να συμμετέχει κάθε στήλη του πίνακα θ.

Πίνακας 4.4: Οι Στήλες του Πίνακα Αναπαράστασης του Μοτίβου και η Συμμετοχή τους στις Κατανομές που Περιλαμβάνουν Θέσεις Μοτίβου.

Στήλη j του θ $(j=1,...,K)$	Κατανομές συμμετοχής $(1,...,2K-1)$	Πλήθος κατανομών
1	$1:K$	K
2	$2:K+1$	K
3	$3:K+2$	K
…	…	…
$K-1$	$K-1:2K-2$	K
K	$K:2K-1$	K

Εφόσον, λοιπόν, ολισθαίνουμε τον πίνακα πάνω σε κάθε υπακολουθία, από αριστερά προς τα δεξιά, αναζητώντας το βέλτιστο ταίριασμα με αυτήν, για $j = 1$ η πρώτη θέση – στήλη του θ ελέγχεται αν συμπίπτει με το τελευταίο σύμβολο της υπακολουθίας. Ανάλογα, για j =2 εξετάζεται αν οι δύο πρώτες θέσεις του θ ταυτίζονται (και σε πόσες το πλήθος θέσεις) με τα δύο τελευταία σύμβολα της υπακολουθίας κ.ο.κ. Τώρα, για $j = K$ γίνεται έλεγχος για το πλήθος των σύμβόλων που είναι κοινά μεταξύ του πίνακα και της υπακολουθίας. Συνεχίζοντας τη μετακίνηση του μοτίβου προς τα αριστερά, για $j = K+1$ εξετάζεται κατά

πόσες θέσεις ταυτίζονται οι θέσεις 2 έως K του θ με τις $K-1$ πρώτες θέσεις (σύμβολα) της υπακολουθίας, για $j=K+2$ ελέγχεται το πλήθος των κοινών θέσεων μεταξύ των στηλών – θέσεων 3 έως K του θ με τις K - 2 πρώτες θέσεις της υπακολουθίας, ... , για $j=2K-1$ εξετάζεται εάν η τελευταία (K - στη) θέση του πίνακα συμπίπτει με το πρώτο σύμβολο της υπακολουθίας. Όπως αντιλαμβάνεται κανείς, από την τελευταία περιγραφή, κάθε στήλη του πίνακα θ συμμετέχει σε K το πλήθος διαφορετικές κατανομές από τις $2K-1$ δυνατές και αυτές αναγράφονται συγκεκριμένα για κάθε στήλη του στη 2η στήλη του Πίνακα 4.4.

Και αυτή η μέθοδος, στοχεύει στη μεγιστοποίηση της λογαριθμικής συνάρτησης Πιθανοφάνειας (με την προϋπόθεση ότι οι υπακολουθίες είναι μεταξύ τους ανεξάρτητες)

$$L(X)=\ln P(X\mid\Psi)=\ln[P(X_1\mid\Psi)\cdot P(X_2\mid\Psi)...P(X_r\mid\Psi)]=\sum_{i=1}^{r}\ln f(X_i\mid\Psi)$$

$$=\sum_{i=1}^{r}\ln\left[\sum_{j=1}^{2K-1}\pi_j\cdot\rho(X_i\mid j,\Theta)+\left(1-\sum_{j=1}^{2K-1}\pi_j\right)\cdot\rho(X_i\mid b)\right].$$

Έχοντας στη διάθεσή μας κάποιες αρχικές τιμές για τις παραμέτρους και θέτοντας $\rho(X_i\mid j,\Theta)=\Phi_j(X_i\mid\Theta)$, $j=1,...,2K-1$ και $\rho(X_i\mid b)=\Phi_{2K}(X_i\mid\Theta)$ σε κάθε επανάληψη του αλγορίθμου, υπολογίζουμε τις εκ των υστέρων πιθανότητες

$$P(Z_i=j\mid X_i,\Psi)=\frac{\pi_j\cdot\Phi_j(X_i\mid\Theta)}{\sum_{l=1}^{2K}\pi_l\cdot\Phi_l(X_i\mid\Theta)},\ j=1,...,2K,\ i=1,...,r \qquad (4.17)$$

και μεγιστοποιώντας τη μέση λογαριθμική πιθανοφάνεια

$$Q=\sum_{i=1}^{r}\sum_{j=1}^{2K}P(Z_i=j\mid X_i,\Psi)\cdot\ln\pi_j+\sum_{i=1}^{r}\sum_{j=1}^{2K}P(Z_i=j\mid X_i,\Psi)\cdot\ln\Phi_j(X_i\mid\Theta) \qquad (4.18)$$

χρησιμοποιώντας τις συνθήκες $\sum_{j=1}^{2K}\pi_j=1$, $\sum_{l=1}^{M}b_l=1$, $\sum_{l=1}^{M}\theta_{lk}=1$, για $k\in\{1,...,K\}$, προκύπτουν οι νέες τιμές για τις παραμέτρους (ανάλογα με αυτές της ενότητας 4.2), από τους τύπους

$$\pi_j = \frac{1}{r}\left(\sum_{i=1}^{r} P(Z_i = j \mid X_i, \Psi)\right),\ j = 1,...,2K, \tag{4.19}$$

$$\theta_{lk} = \frac{\sum_{i=1}^{r}\left[\sum_{j=k}^{K+k-1} P(Z_i = j \mid X_i, \Psi)\cdot \delta(X_{i_{K-j+k}}, a_l)\right]}{\sum_{i=1}^{r}\left[\sum_{j=k}^{K+k-1} P(Z_i = j \mid X_i, \Psi)\right]},\ \ l = 1,...,M,\ \ k = 1,...,K, \tag{4.20}$$

$$b_l = \frac{\sum_{i=1}^{r}\left[\sum_{j=1}^{K} P(Z_i = j \mid X_i, \Psi)\cdot \sum_{k=1}^{K-j}\delta(X_{i_k}, a_l) + \sum_{j=K+1}^{2K} P(Z_i = j \mid X_i, \Psi)\cdot \sum_{k=2K-j+1}^{K}\delta(X_{i_k}, a_l)\right]}{\sum_{i=1}^{r}\left[\sum_{j=1}^{K}(K-j)\cdot P(Z_i = j \mid X_i, \Psi) + \sum_{j=K+1}^{2K}(j-K)\cdot P(Z_i = j \mid X_i, \Psi)\right]}, \tag{4.21}$$

$l = 1,...,M$.

Η σχέση (4.21) προέκυψε από το γεγονός ότι κάθε μία από τις κατανομές $\Phi_j,\ \ j \in \{1,...K-1, K+1,...,2K\}$, αφορά υπακολουθίες, οι οποίες περιλαμβάνουν και ένα πλήθος θέσεων background (από μία μέχρι $K-1$). Έτσι, ο αλγόριθμος έχει, τώρα, ως εξής

Πίνακας 4.5: Ο Αλγόριθμος του Οριοθετούμενου Πίνακα (Bounded Array).

- Αρχικά: Θεώρησε τυχαίες τιμές για τις παραμέτρους $\Psi^{(\pi)}$.
- Επανάλαβε τα παρακάτω
  - <u>*Expectation – βήμα:*</u> Υπολόγισε τις εκ των υστέρων πιθανότητες

$$P(Z_i = j \mid X_i, \Psi) = \frac{\pi_j \cdot \Phi_j(X_i \mid \Theta)}{\sum_{l=1}^{2K}\pi_l \cdot \Phi_l(X_i \mid \Theta)},\ i = 1,...,r,\ j = 1,2,...,2K.$$

  - <u>*Maximization – βήμα:*</u> Εκτίμησε τις νέες παραμέτρους $\Psi^{(\nu)} = \arg\max_{\Psi} Q(\Psi, \Psi^{(\pi)})$,

όπου

$$Q(\Psi,\Psi^{(\pi)})=\sum_{i=1}^{r}\sum_{j=1}^{2K}P(Z_i=j\mid X_i,\Psi)\cdot\ln P(X_i,Z_i=j\mid\Psi),$$

από τις σχέσεις:

1) $\pi_j=\frac{1}{r}\left(\sum_{i=1}^{r}P(Z_i=j\mid X_i,\Psi)\right),\ j=1,2,...,2K$

2) $\theta_{lk}=\frac{\sum_{i=1}^{r}\left[\sum_{j=k}^{K+k-1}P(Z_i=j\mid X_i,\Psi)\cdot\delta(X_{i_{K-j+k}},a_l)\right]}{\sum_{i=1}^{r}\left[\sum_{j=k}^{K+k-1}P(Z_i=j\mid X_i,\Psi)\right]},\ l=1,...,M,\ k=1,...,K$

3) $b_l=\frac{\sum_{i=1}^{r}\left[\sum_{j=1}^{K}P(Z_i=j\mid X_i,\Psi)\cdot\sum_{k=1}^{K-j}\delta(X_{i_k},a_l)+\sum_{j=K+1}^{2K}P(Z_i=j\mid X_i,\Psi)\cdot\sum_{k=2K-j+1}^{K}\delta(X_{i_k},a_l)\right]}{\sum_{i=1}^{r}\left[\sum_{j=1}^{K}(K-j)\cdot P(Z_i=j\mid X_i,\Psi)+\sum_{j=K+1}^{2K}(j-K)\cdot P(Z_i=j\mid X_i,\Psi)\right]},$

$l=1,...,M$,

όσο ισχύει $Q(\Psi,\Psi^{(\nu)})>Q(\Psi,\Psi^{(\pi)})+\varepsilon$, όπου $\varepsilon>0$.

Όταν ολοκληρωθούν οι επαναλήψεις του αλγορίθμου και λάβουμε τις τελικές τιμές των παραμέτρων, για να συμπεράνουμε ποιες θέσεις από τον πίνακα θ αντιστοιχούν σε θέσεις μοτίβου, ενεργούμε ως εξής. Υπολογίζουμε το μέγιστο στοιχείο $\left(\max_i\right)_{i=1}^{K}$ κάθε στήλης του θ και στη συνέχεια, βρίσκουμε για ποιες στήλες (για ποια $i\in\{1,...,K\}$) ισχύει η συνθήκη $\max_i\geq 0.4$. Στη συνέχεια, κρατάμε την πρώτη και την τελευταία από τις θέσεις αυτές, οι οποίες αντιστοιχούν στην πρώτη και την τελευταία στήλη του θ που αντιστοιχούν σε θέσεις μοτίβου. Έτσι, ανάλογα με το πλήθος των θέσεων μοτίβου που ανακαλύφθηκαν, καθώς και τη θέση τους μέσα στον θ, επιλέγουμε και την κατανομή από την οποία θα εξάγουμε τις υπακολουθίες που αποτελούν μοτίβο και πάλι, με βάση τις εκ των υστέρων πιθανότητες (όπως δηλαδή στον ΕΜ και στην πρώτη εναλλακτική μορφή του). Για να γίνει κατανοητό

πώς ακριβώς γίνεται η επιλογή της κατανομής, παρατίθεται ακολούθως ένα γενικό παράδειγμα που περιλαμβάνει όλες τις δυνατές περιπτώσεις κατανομών (για το μοτίβο).

Μια περίπτωση είναι αυτή, όπου οι στήλες του θ που ικανοποιούν τον περιορισμό $\max_i \geq 0.4$ εκτείνονται από την $1^{η}$ μέχρι την $t-$στη, δηλαδή $t \in \{1,...,K-1\}$. Τότε, ως κατανομή του μοτίβου επιλέγεται η Φ_t.

Αν, τώρα, οι ζητούμενες στήλες του θ είναι η $t-$στη (για $t>1$) μέχρι την $K-$στη, επιλέγουμε τη Φ_{K+t-1} ως την κατανομή που αντιστοιχεί στο μοτίβο.

Ασφαλώς, αν οι ζητούμενες θέσεις αντιστοιχούν στις στήλες 1 έως K του θ, τότε ο πίνακας θ αναπαριστά ολόκληρο το μοτίβο και η ζητούμενη κατανομή είναι η Φ_K.

Τέλος, εάν οι στήλες που μας ενδιαφέρουν μέσα στον θ είναι από την t_1-στη έως την t_2-στη με $1<t_1 \leq t_2 < K$, ανάλογα με το ποιο στοιχείο από τα $\max_{t_1}, \max_{t_2}$ είναι το μεγαλύτερο, επιλέγουμε είτε την κατανομή Φ_{K+t_1-1} ή τη Φ_{t_2} αντίστοιχα.

ΚΕΦΑΛΑΙΟ 5. ΠΕΙΡΑΜΑΤΙΚΗ ΜΕΛΕΤΗ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ ΕΝΤΟΠΙΣΜΟΥ ΜΟΤΙΒΩΝ ΣΕ ΣΥΜΒΟΛΟΣΕΙΡΕΣ

5.1. Πειραματική Διαδικασία

Η πειραματική μελέτη των αλγορίθμων που περιγράψαμε στις προηγούμενες ενότητες διεξήχθη τόσο σε πειραματικά περιβάλλοντα με τεχνητά δεδομένα όσο και σε πραγματικά δεδομένα. Στόχος αυτής της διαδικασίας είναι να μετρήσουμε τις επιδόσεις των μεθόδων και να διαπιστώσουμε κατά πόσο οι ήδη υπάρχουσες, καθώς και οι προτεινόμενες μπορούν να προσφέρουν βέλτιστες λύσεις στο πρόβλημα του εντοπισμού μοτίβων σε συμβολοσειρές.

Καταρχήν, για τη δημιουργία των *τεχνητών δεδομένων* ακολουθήσαμε την παρακάτω διαδικασία. Κατασκευάσαμε αλυσίδες συμβόλων, μεταβλητού μήκους (ογδόντα (80) έως εκατό (100) συμβόλων), από ένα καθορισμένο αλφάβητο A, με ομοιόμορφο τρόπο. Συγκεκριμένα, για κάθε θέση τους, επιλέγεται, με τυχαίο τρόπο ένα από τα σύμβολα του A. Δοθέντος, τώρα, ενός μοτίβου συγκεκριμένου μήκους K (με χαρακτήρες από το A), τοποθετήσαμε σε μία ορισμένη θέση κάθε συμβολοσειράς (η οποία επιλέχθηκε τυχαία, λαμβάνοντας υπόψη το μήκος της) ένα μεταλλαγμένο αντίγραφό του, με βάση μια πιθανότητα μετάλλαξης $p_{mut} \in [0,1]$. Έτσι, με βάση τα «μεταλλαγμένα» μοτίβα που εισήχθηκαν μέσα στις ακολουθίες, υπολογίσαμε τον πραγματικό πίνακα βάρους θέσης του

μοτίβου θ_{true}. Στη συνέχεια, κρατώντας τη θέση κάθε αλυσίδας, όπου τοποθετήθηκε το μοτίβο, δημιουργήσαμε από το σύνολο όλων των συμβολοσειρών, όλες τις δυνατές υπακολουθίες $X = \{X_1, X_2, ..., X_r\}$ μήκους K που προκύπτουν από αυτές. Το σύνολο X είναι αυτό που χρησιμοποιήθηκε ως το σύνολο εκπαίδευσης του προβλήματος και με βάση το οποίο, οι μέθοδοι που εφαρμόσαμε προσπάθησαν να προσδιορίσουν το κρυμμένο μοτίβο.

Συγκεκριμένα, σε κάθε τέτοιο σύνολο εφαρμόσαμε, αρχικά, τον συνθετικό αλγόριθμο ομαδοποίησης (Agglomerative), προκειμένου να λάβουμε μια καλή αρχικοποίηση της παραμέτρου που μας ενδιαφέρει (δηλαδή του πίνακα βάρους θέσης θ του μοτίβου) και αρχικοποιήσαμε με τυχαίο τρόπο τις υπόλοιπες παραμέτρους. Αυτές, λοιπόν, δόθηκαν ως είσοδοι στους αλγορίθμους ΕΜ, Εκτεταμένου και Οριοθετούμενου πίνακα (όπως περιγράφηκαν στο 4^{ο} κεφάλαιο) και στον αλγόριθμο Δειγματοληψίας του Gibbs (Gibbs Sampling). Έπειτα, γνωρίζοντας ότι ο αλγόριθμος των Κ-κέντρων (K-means) δεν παράγει πάντα μια καλή αρχικοποίηση για τον θ, τον εφαρμόσαμε στο ίδιο σύνολο δεδομένων X είκοσι φορές και δώσαμε την καλύτερη δυνατή αρχικοποίηση από αυτές (κατόπιν συγκρίσεώς της με τον θ_{true}) για την παράμετρο ενδιαφέροντος. Οι υπόλοιπες παράμετροι αρχικοποιήθηκαν, και πάλι, τυχαία και όλο το σύνολο των παραμέτρων δόθηκε ως είσοδος στους ίδιους αλγορίθμους (ΕΜ, Extended Array, Bounded Array και Gibbs Sampling).

Βασικός στόχος των παραπάνω πειραμάτων είναι, προφανώς, η σύγκριση μεταξύ των δύο αλγορίθμων ομαδοποίησης, όσον αφορά την αρχικοποίηση που παράγουν και ασφαλώς, μεταξύ των τεσσάρων μεθόδων που αποσκοπούν στην ανίχνευση των αντιγράφων του μοτίβου μέσα στις συμβολοσειρές.

Προκειμένου, τώρα, να μελετήσουμε τις επιδόσεις των αλγορίθμων, χρησιμοποιήσαμε τρία είδη μέτρων και συγκεκριμένα, τις ποσότητες $\Delta\theta$, Sn και Sp. Οι δύο τελευταίες αφορούν τα ποσοστά επιτυχίας των αλγορίθμων ως προς τον εντοπισμό των αντιγράφων του μοτίβου μέσα στις αλυσίδες.

Αναλυτικά, καθένα από τα παραπάνω μέτρα εκφράζουν τα εξής:

- $\Delta\theta$: το άθροισμα των απολύτων διαφορών των αντίστοιχων στοιχείων του πίνακα βάρους θέσης θ του μοτίβου που παράγει κάθε μία από τις τέσσερις βασικές μεθόδους και του πίνακα θ_{true} (δηλαδή η L_1-νόρμα αν θεωρήσουμε τον πίνακα $\theta_{true}-\theta$ ως ένα διάνυσμα διάστασης $1xM\cdot K$).
- Sn: το ποσοστό των θέσεων που βρίσκει ο εκάστοτε αλγόριθμος από όλα τα αντίγραφα που έχουν τοποθετηθεί μέσα στις συμβολοσειρές. Μάλιστα, αν έχουν βρεθεί δύο επικαλύψεις από το αντίγραφο του μοτίβου σε μια αλυσίδα, ως ποσοστό επιτυχίας θεωρείται ο μέσος όρος των ποσοστών επιτυχίας αυτών των δύο και εφόσον γνωρίζουμε ότι έχουμε εισάγει N το πλήθος αντίγραφα του μοτίβου (αφού υπάρχει ένα σε κάθε συμβολοσειρά), προφανώς υπάρχει ένα αντίγραφο (σε κάποια από τις N αλυσίδες), για το οποίο σημειώνεται μηδενικό ποσοστό (επιτυχίας). Για αυτόν τον λόγο, κρατάμε τη θέση εμφάνισης του μοτίβου σε κάθε αλυσίδα και επιπλέον, σε ποια αλυσίδα και σε ποια θέση της εμφανίζεται κάθε υπακολουθία μήκους K .
- Sp: το ποσοστό των αντιγράφων εκείνων, από τα οποία βρέθηκε τουλάχιστον το 33% των συνολικών τους θέσεων. Και αυτό υπολογίζεται με ανάλογο τρόπο (δηλαδή λαμβάνοντας υπόψη όλες τις επικαλύψεις για κάθε αντίγραφο) όπως και το Sn .

Στα πειράματα που πραγματοποιήθηκαν σε τυχαία δεδομένα, υπολογίσαμε όλα τα παραπάνω μέτρα για διαφορετικό πλήθος συμβολοσειρών ($N=10,\ 15,\ 20$), συμβόλων του αλφάβητου ($|A|=10,\ 12,\ldots,\ 20$), μήκος ($K=8,\ 12,\ 16$) και είδος μοτίβου (επαναληπτικό, όπως LALALALA ή όχι), καθώς και διαφορετική πιθανότητα μετάλλαξης ($p_{mut}=.05,\ .1,\ \ldots,$ .4). Τα αποτελέσματα που προέκυψαν θα παρουσιασθούν υπό τη μορφή γραφημάτων, ώστε να γίνεται εύκολα αντιληπτή η απόδοση της κάθε μεθόδου σε κάθε μία από τις παραπάνω περιπτώσεις.

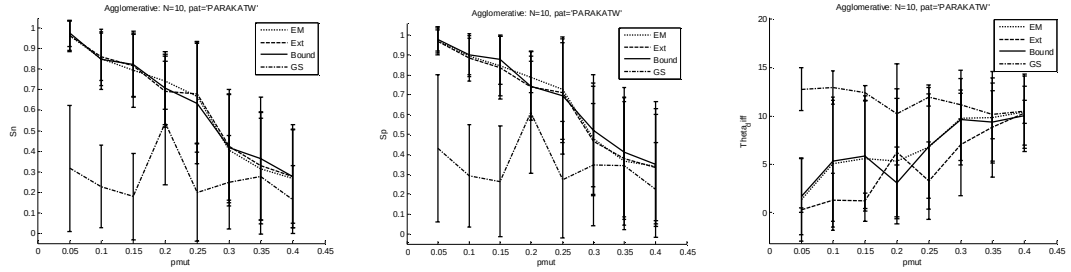
Τέλος, όσον αφορά τα πραγματικά δεδομένα, αυτά αφορούν αλυσίδες DNA (δηλαδή αλφάβητο είναι το $A=\{A,G,C,T\}$) ίδιου μήκους, όπου το μοτίβο ξεκινάει είτε στην 21^{η} είτε στην 51^{η} θέση κάθε συμβολικής ακολουθίας. Και στην περίπτωση αυτή, ακολουθείται η ίδια ακριβώς διαδικασία εφαρμογής των βασικών μεθόδων και υπολογίζονται τα ίδια ποσοστά απόδοσης.

5.2. Γραφικές Παραστάσεις

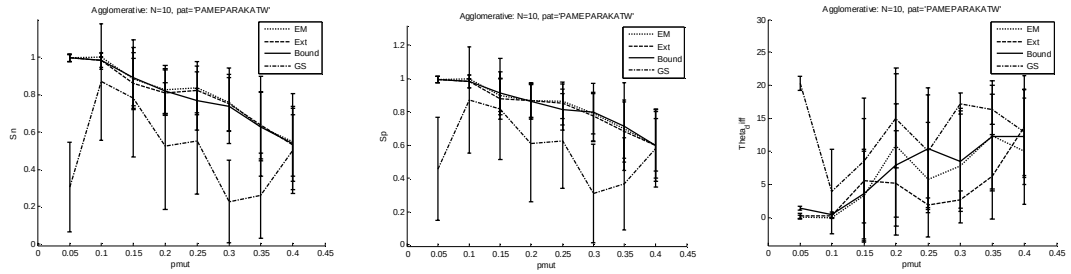
Στις ενότητες που ακολουθούν, παρατίθενται τα γραφήματα που προέκυψαν από τα πειράματα που πραγματοποιήσαμε, καθώς και τα συμπεράσματα που εξάγονται από αυτά κατόπιν παρατηρήσεώς τους. Σε ό,τι ακολουθεί, με Ext (Extended array) συμβολίζουμε την $1^{η}$ παραλλαγή του αλγορίθμου EM (ή μέθοδο του Εκτεταμένου πίνακα), με Bound (Bounded array) την $2^{η}$ παραλλαγή του EM (ή μέθοδο του Οριοθετούμενου πίνακα) και με GS την μέθοδο Δειγματοληψίας του Gibbs.

5.2.1. Μη Επαναληπτικό Μοτίβο και Συνθετικός Αλγόριθμος

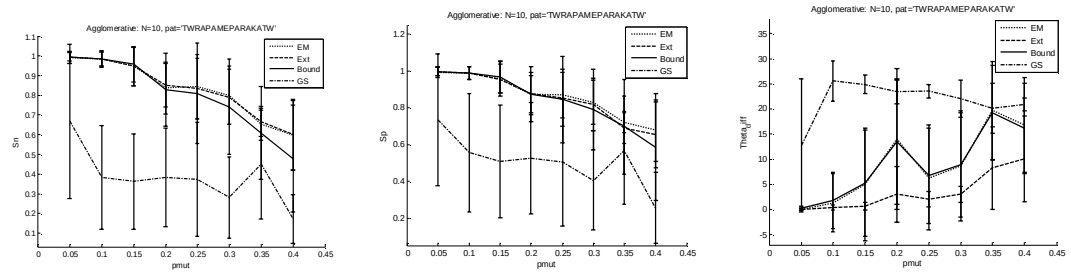
Αρχικά, θα παρουσιάσουμε τα γραφήματα που προέκυψαν στην περίπτωση του μη επαναληπτικού μοτίβου με αρχικοποίηση του συνθετικού αλγόριθμου (Agglomerative). Συγκεκριμένα, χρησιμοποιήσαμε τα ακόλουθα μοτίβα διαφορετικού μήκους: $m_1 = ['PARAKATW']$ (για $K = 8$), $m_2 = ['PAMEPARAKATW']$ (για $K = 12$), $m_2 = ['TWRAPAMEPARAKATW']$ (για $K = 16$). Χρησιμοποιώντας, λοιπόν, $N = 10,15$ και 20 συμβολοσειρές, προέκυψαν τα γραφήματα που απεικονίζονται στα σχήματα 5.1 – 5.9.



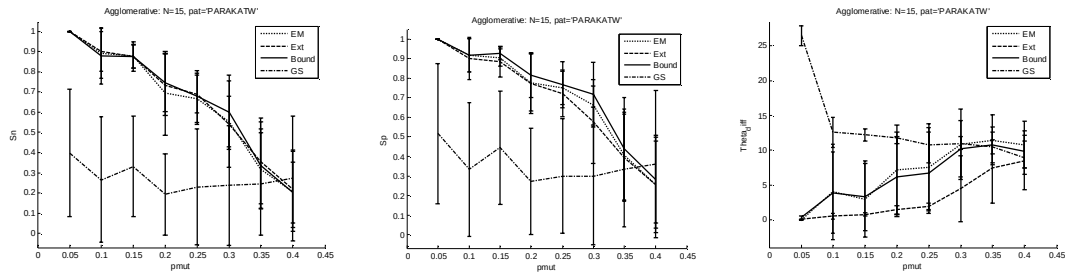
Σχήμα 5.1: Μέτρα επίδοσης για μη επαναληπτικό μοτίβο μήκους $K = 8$ συμβόλων μέσα σε $N = 10$ συμβολικές ακολουθίες και χρήση του συνθετικού αλγόριθμου ομαδοποίησης.



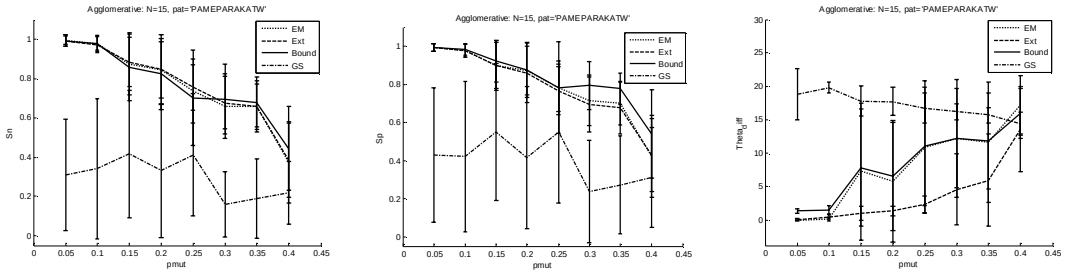
Σχήμα 5.2: Μέτρα επίδοσης για μη επαναληπτικό μοτίβο μήκους $K = 12$ συμβόλων μέσα σε $N = 10$ συμβολικές ακολουθίες και χρήση του συνθετικού αλγόριθμου ομαδοποίησης.



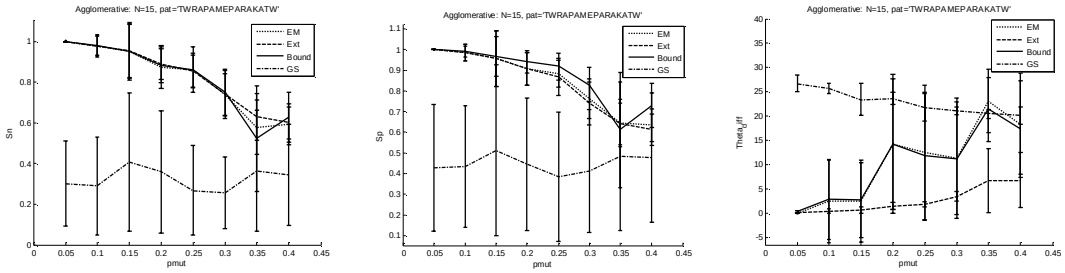
Σχήμα 5.3: Μέτρα επίδοσης για μη επαναληπτικό μοτίβο μήκους $K = 16$ συμβόλων μέσα σε $N = 10$ συμβολικές ακολουθίες και χρήση του συνθετικού αλγόριθμου ομαδοποίησης.



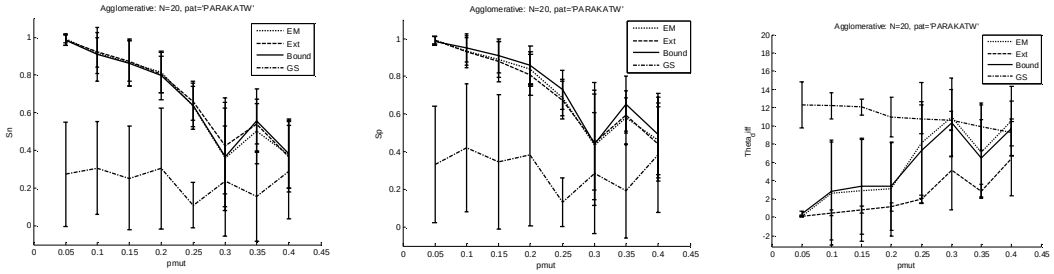
Σχήμα 5.4: Μέτρα επίδοσης για μη επαναληπτικό μοτίβο μήκους $K = 8$ συμβόλων μέσα σε $N = 15$ συμβολικές ακολουθίες και χρήση του συνθετικού αλγόριθμου ομαδοποίησης.



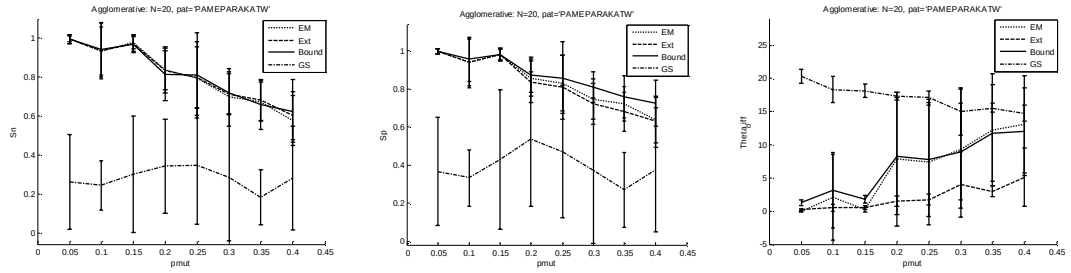
Σχήμα 5.5: Μέτρα επίδοσης για μη επαναληπτικό μοτίβο μήκους $K = 12$ συμβόλων μέσα σε $N = 15$ συμβολικές ακολουθίες και χρήση του συνθετικού αλγόριθμου ομαδοποίησης.



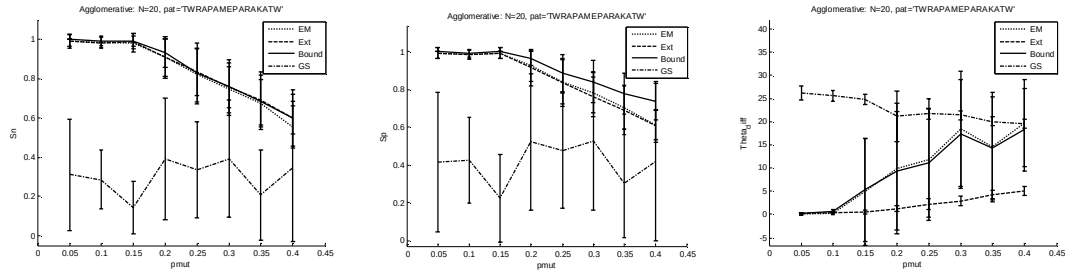
Σχήμα 5.6: Μέτρα επίδοσης για μη επαναληπτικό μοτίβο μήκους $K = 16$ συμβόλων μέσα σε $N = 15$ συμβολικές ακολουθίες και χρήση του συνθετικού αλγόριθμου ομαδοποίησης.



Σχήμα 5.7: Μέτρα επίδοσης για μη επαναληπτικό μοτίβο μήκους $K = 8$ συμβόλων μέσα σε $N = 20$ συμβολικές ακολουθίες και χρήση του συνθετικού αλγόριθμου ομαδοποίησης.



Σχήμα 5.8: Μέτρα επίδοσης για μη επαναληπτικό μοτίβο μήκους $K = 12$ συμβόλων μέσα σε $N = 20$ συμβολικές ακολουθίες και χρήση του συνθετικού αλγόριθμου ομαδοποίησης.



Σχήμα 5.9: Μέτρα επίδοσης για μη επαναληπτικό μοτίβο μήκους $K = 16$ συμβόλων μέσα σε $N = 20$ συμβολικές ακολουθίες και χρήση του συνθετικού αλγόριθμου ομαδοποίησης.

Όπως παρατηρούμε, αρχικά, για $N = 10$, καθώς η πιθανότητα μετάλλαξης (*pmut*) αυξάνει (οριζόντιος άξονας) τα μέτρα *Sn* και *Sp* (κατακόρυφος άξονας) παρουσιάζουν μια σταθερή πτώση, ενώ η ποσότητα $\Delta\theta$ αυξάνεται. Για μικρές τιμές της πιθανότητας μετάλλαξης δεν καθίσταται εμφανές ποια από τις τρεις μεθόδους EM, Ext και Bound υπερτερεί όσον αφορά τα ποσοστά *Sn* και *Sp*, ενώ καθώς αυτή αυξάνεται, αρχίζει να παρατηρείται μια μικρή διαφορά μεταξύ των τριών μεθόδων. Παρ' όλα αυτά, την χαμηλότερη τιμή $\Delta\theta$ (και συνεπώς, την καλύτερη προσέγγιση του μοτίβου) εμφανίζει η μέθοδος του εκτεταμένου πίνακα, με εξαίρεση τις τιμές $p_{mut} = 0.2$ και 0.15, 0.4 για $K = 8$ και 12 αντίστοιχα, όπου υπερέχει η μέθοδος του οριοθετούμενου πίνακα (Bound). Ο Δειγματολήπτης Gibbs εμφανίζει πολύ χαμηλότερες τιμές στις περιπτώσεις, όπου το μοτίβο έχει μήκος $K = 8, 12$ και όπως είναι αναμενόμενο, παρουσιάζει τις υψηλότερες τιμές για την ποσότητα $\Delta\theta$.

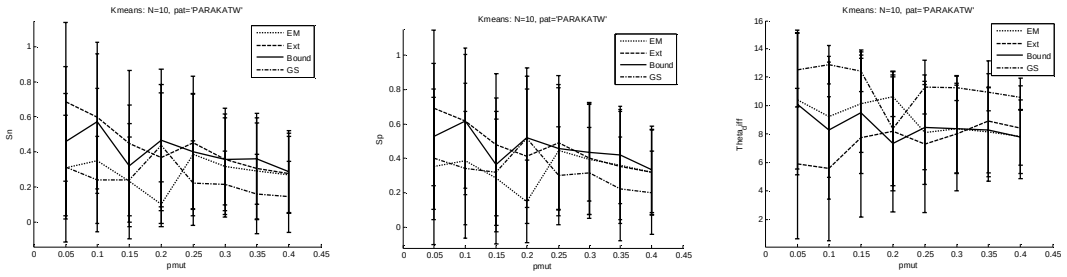
Επίσης, πιο απότομη πτώση στα Sn και Sp παρατηρείται για την τιμή $pmut = 0.3$ όταν $K = 8$, ενώ οι δύο νέες μέθοδοι υπερέχουν στη γενική περίπτωση έναντι του EM. Πιο συγκεκριμένα, για το μικρότερο μήκος μοτίβου $K = 8$, υπερτερεί η μέθοδος Bound (του οριοθετούμενου πίνακα), ενώ καθώς αυτό αυξάνεται η Ext παρουσιάζει μεγαλύτερη ευστάθεια (δηλαδή έχει τις καλύτερες επιδόσεις).

Η συμπεριφορά των τεσσάρων μεθόδων δεν αποκλίνει σημαντικά από την προηγούμενη περιγραφή, καθώς το μήκος του μοτίβου αυξάνει, με τη διαφορά ότι η μείωση στα μέτρα Sn και Sp είναι μικρότερη (καθώς αυξάνει η πιθανότητα μετάλλαξης).

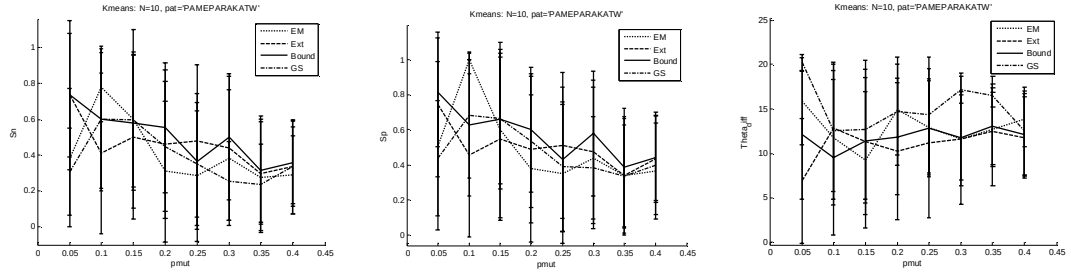
Καθώς, τώρα, αυξάνει το πλήθος των συμβολοσειρών, η γενική συμπεριφορά των τεσσάρων μεθόδων δεν αλλάζει σημαντικά. Εν τούτοις, στην περίπτωση που είναι $N = 15$, αξίζει να παρατηρήσουμε ότι η μέθοδος Δειγματοληψίας του Gibbs εμφανίζει, για $K = 8$, μια σταθερή άνοδο ως προς τα μέτρα Sn και Sp και πτώση της ποσότητας $\Delta\theta$, με αποτέλεσμα για την μέγιστη πιθανότητα μετάλλαξης $p_{mut} = 0.4$ να παρουσιάζει τα υψηλότερα ποσοστά από όλες τις υπόλοιπες μεθόδους στα Sn και Sp. Επιπλέον, για $K = 12$ και 16, η Bound εμφανίζει τις υψηλότερες τιμές στο ποσοστό Sp, ενώ παρόμοια αποτελέσματα προκύπτουν και όταν είναι $N = 20$.

5.2.2. Μη Επαναληπτικό Μοτίβο και Αρχικοποίηση των K-κέντρων

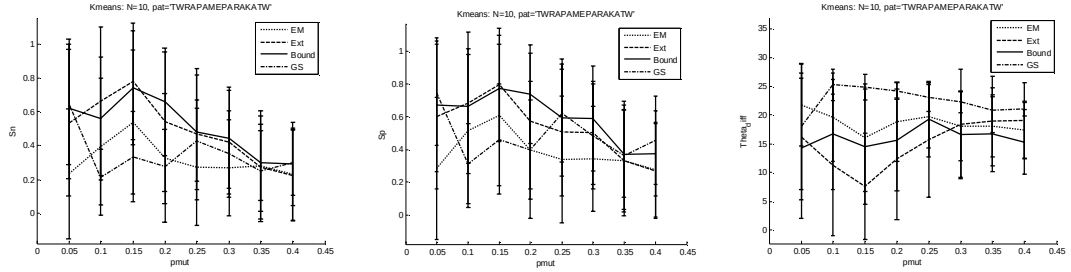
Εν συνεχεία, στα σχήματα 5.10 – 5.18 απεικονίζονται τα γραφήματα που προέκυψαν από τα πειράματα χρησιμοποιώντας αρχικοποίηση του αλγορίθμου των K-κέντρων (K-means), όπως περιγράφηκε στην παράγραφο 5.1, για τα ίδια μήκη μοτίβων ($K = 8, 12, 16$) και το ίδιο πλήθος συμβολοσειρών ($N = 10,\ 15, 20$).



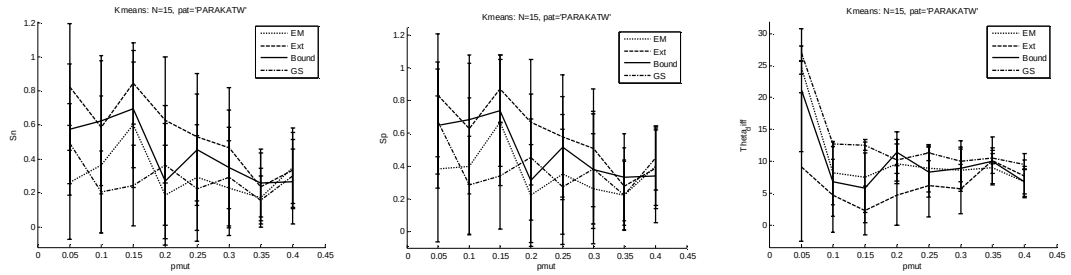
Σχήμα 5.10: Μέτρα επίδοσης για μη επαναληπτικό μοτίβο μήκους $K = 8$ συμβόλων μέσα σε $N = 10$ συμβολικές ακολουθίες και χρήση του αλγορίθμου των K-κέντρων.



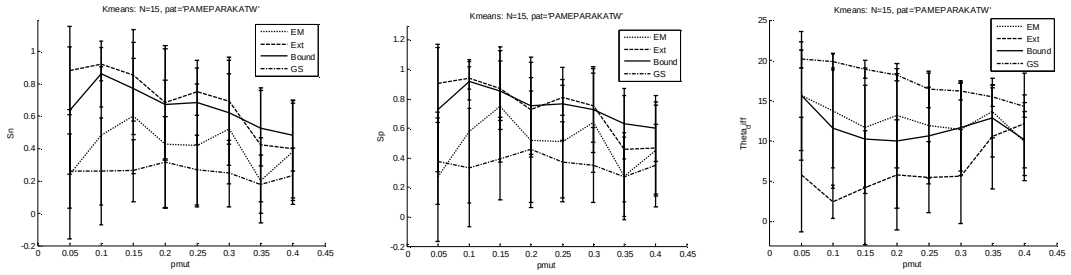
Σχήμα 5.11: Μέτρα επίδοσης για μη επαναληπτικό μοτίβο μήκους $K = 12$ συμβόλων μέσα σε $N = 10$ συμβολικές ακολουθίες και χρήση του αλγορίθμου των K-κέντρων.



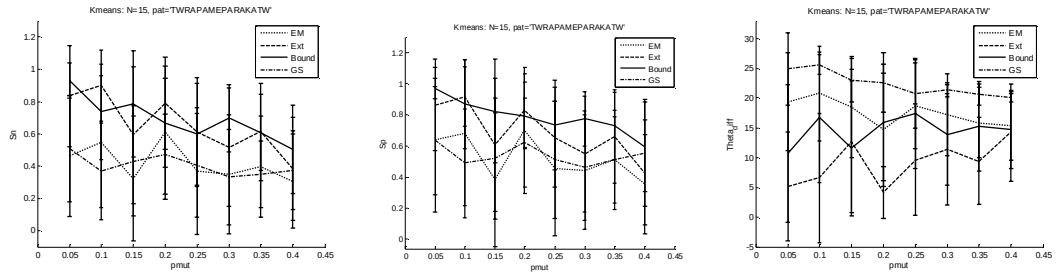
Σχήμα 5.12: Μέτρα επίδοσης για μη επαναληπτικό μοτίβο μήκους $K = 16$ συμβόλων μέσα σε $N = 10$ συμβολικές ακολουθίες και χρήση του αλγορίθμου των K-κέντρων.



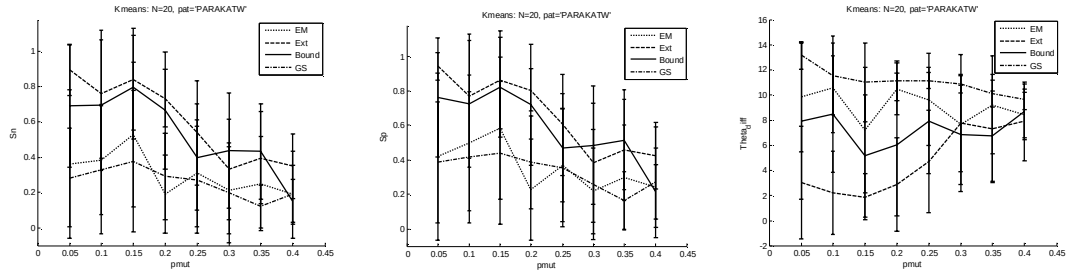
Σχήμα 5.13: Μέτρα επίδοσης για μη επαναληπτικό μοτίβο μήκους $K = 8$ συμβόλων μέσα σε $N = 15$ συμβολικές ακολουθίες και χρήση του αλγορίθμου των K-κέντρων.



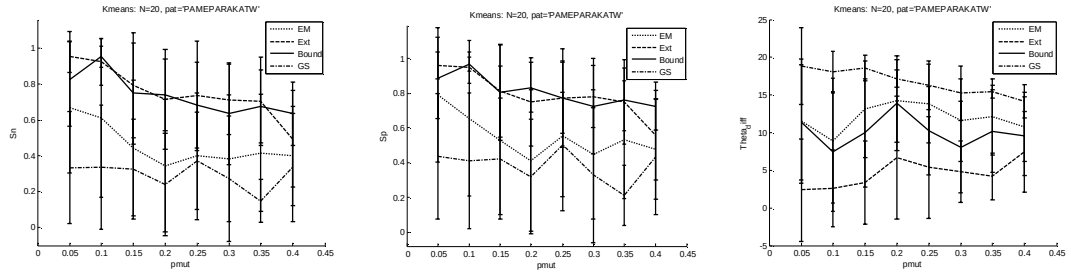
Σχήμα 5.14: Μέτρα επίδοσης για μη επαναληπτικό μοτίβο μήκους $K = 12$ συμβόλων μέσα σε $N = 15$ συμβολικές ακολουθίες και χρήση του αλγορίθμου των K-κέντρων.



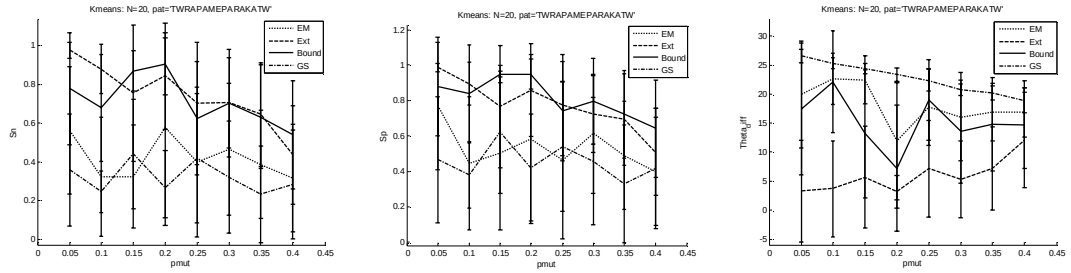
Σχήμα 5.15: Μέτρα επίδοσης για μη επαναληπτικό μοτίβο μήκους $K = 16$ συμβόλων μέσα σε $N = 15$ συμβολικές ακολουθίες και χρήση του αλγορίθμου των K-κέντρων.



Σχήμα 5.16: Μέτρα επίδοσης για μη επαναληπτικό μοτίβο μήκους $K = 8$ συμβόλων μέσα σε $N = 20$ συμβολικές ακολουθίες και χρήση του αλγορίθμου των K-κέντρων.



Σχήμα 5.17: Μέτρα επίδοσης για μη επαναληπτικό μοτίβο μήκους $K = 12$ συμβόλων μέσα σε $N = 20$ συμβολικές ακολουθίες και χρήση του αλγορίθμου των K-κέντρων.



Σχήμα 5.18: Μέτρα επίδοσης για μη επαναληπτικό μοτίβο μήκους $K = 16$ συμβόλων μέσα σε $N = 20$ συμβολικές ακολουθίες και χρήση του αλγορίθμου των K-κέντρων.

Καταρχήν, όπως είναι φανερό εδώ, λόγω του ότι η αρχικοποίηση του πίνακα θ δεν είναι πάντοτε τόσο καλή όσο αυτή που παρέχεται από τον συνθετικό αλγόριθμο ομαδοποίησης, η αύξηση της πιθανότητας μετάλλαξης δεν συνεπάγεται και μείωση των μέτρων απόδοσης Sn και Sp (όπως συνέβαινε προηγουμένως). Κάτι που αξίζει, ακόμα, να παρατηρήσει κανείς είναι ότι για μικρή τιμή της παραμέτρου μετάλλαξης p_{mut}, η τυπική απόκλιση είναι σημαντική στην περίπτωση της αρχικοποίησης με τον αλγόριθμο των K-κέντρων συγκρινόμενη με αυτή που παρέχεται στην αρχικοποίηση του συνθετικού αλγορίθμου, καθώς ο δεύτερος δίνει μια πολύ ικανοποιητική προσέγγιση του πίνακα περιγραφής του μοτίβου.

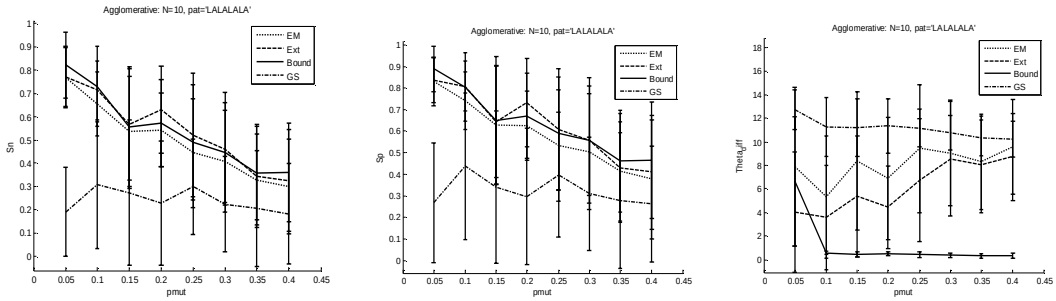
Επίσης, εξαιρετικά ενδιαφέρον είναι το γεγονός ότι ο αλγόριθμος EM υστερεί σημαντικά σε σύγκριση με τις δύο νέες μεθόδους όσον αφορά και τα τρία μέτρα επίδοσης, όταν δεν υπάρχει καλή αρχικοποίηση του παράγοντα ενδιαφέροντος (πίνακα θ). Μάλιστα, για $K = 12$ και $p_{mut} = 0.4$, ακόμα και η μέθοδος Δειγματοληψίας του Gibbs δίνει πολύ καλύτερες τιμές

για τα Sn, Sp και $\Delta\theta$ από ότι ο EM. Το ίδιο συμβαίνει με τις ποσότητες Sn και Sp για το μέγιστο μήκος μοτίβου που εξετάσαμε ($K=16$), καθώς είναι πολύ μεγαλύτερη η πιθανότητα να το εντοπίσει ο Δειγματολήπτης Gibbs, ακόμα και με την τυχαία διαδικασία επιλογής που χρησιμοποιεί. Έτσι, για $p_{mut}=0.05$ υπερέχει έναντι όλων των υπόλοιπων μεθόδων και μάλιστα, είναι εμφανές ότι δίνει και μικρότερη τιμή $\Delta\theta$ έναντι του EM, για $p_{mut}=0.25,0.3$ παρουσιάζει υψηλότερα ποσοστά Sn και Sp σε σχέση με τον ίδιο αλγόριθμο και επικρατεί έναντι όλων στο μέτρο Sp για $p_{mut}=0.3,0.4$. Βέβαια, δεν μπορεί να συγκριθεί με τις δύο νέες μεθόδους όσον αφορά την προσέγγιση του μοτίβου (ποσότητα $\Delta\theta$).

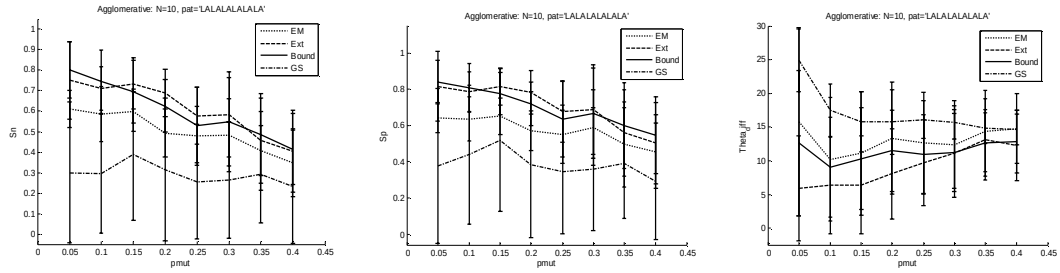
Τέλος, στην προκειμένη περίπτωση, όπως εύκολα μπορεί να διακρίνει κανείς από τα σχήματα 5.13 –5.18, τα αποτελέσματα δεν διαφοροποιούνται με την αύξηση του πλήθους των συμβολοσειρών.

5.2.3. Επαναληπτικό Μοτίβο και Συνθετικός Αλγόριθμος

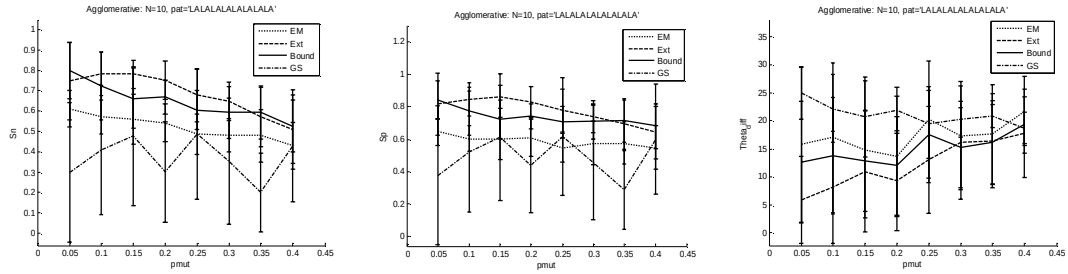
Στην περίπτωση του επαναληπτικού μοτίβου, παραστήσαμε γραφικά τις επιδόσεις των τεσσάρων μεθόδων, για τα μοτίβα $m_1=['LALALALA']$ $(K=8)$, $m_2=['LALALALALALA']$ $(K=12)$ και $m_3=['LALALALALALALALA']$ $(K=16)$ αντίστοιχα και πάλι, για $N=10,\ 15$ και 20 το πλήθος συμβολοσειρές.



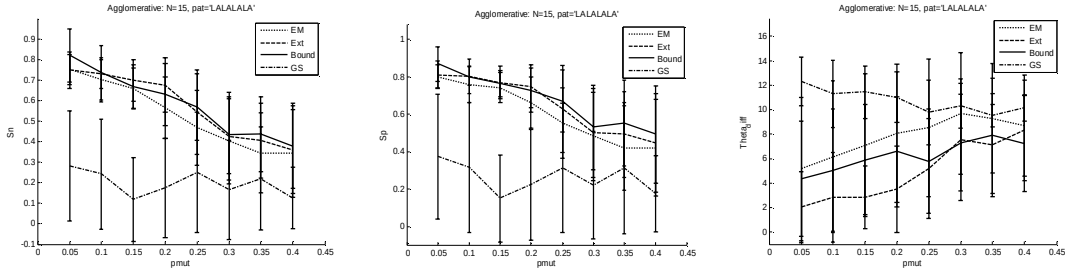
Σχήμα 5.19: Μέτρα επίδοσης για επαναληπτικό μοτίβο μήκους $K=8$ συμβόλων μέσα σε $N=10$ συμβολικές ακολουθίες και χρήση του συνθετικού αλγόριθμου.



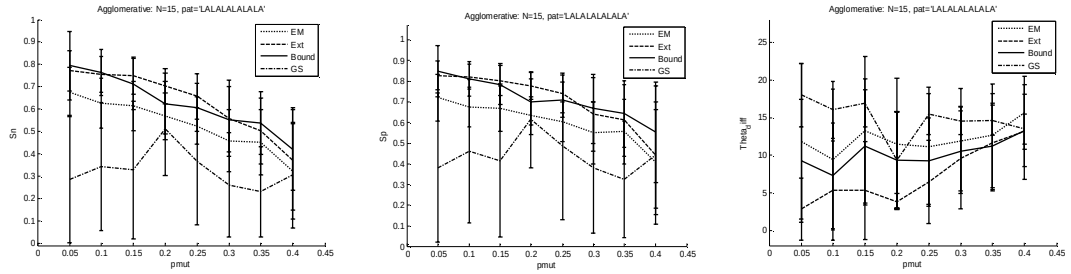
Σχήμα 5.20: Μέτρα επίδοσης για επαναληπτικό μοτίβο μήκους $K = 12$ συμβόλων μέσα σε $N = 10$ συμβολικές ακολουθίες και χρήση του συνθετικού αλγόριθμου.



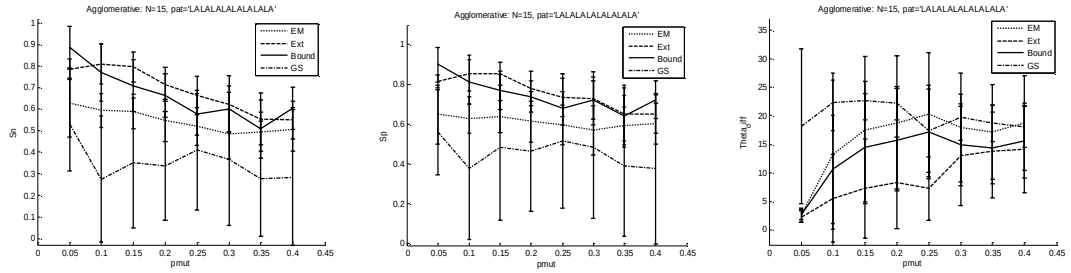
Σχήμα 5.21: Μέτρα επίδοσης για επαναληπτικό μοτίβο μήκους $K = 16$ συμβόλων μέσα σε $N = 10$ συμβολικές ακολουθίες και χρήση του συνθετικού αλγόριθμου.



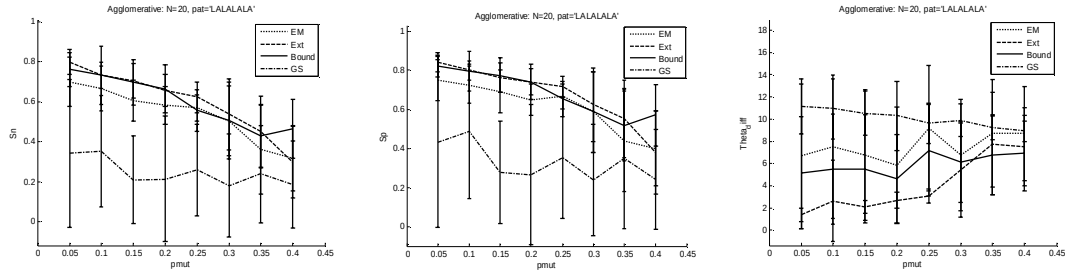
Σχήμα 5.22: Μέτρα επίδοσης για επαναληπτικό μοτίβο μήκους $K = 8$ συμβόλων μέσα σε $N = 15$ συμβολικές ακολουθίες και χρήση του συνθετικού αλγόριθμου.



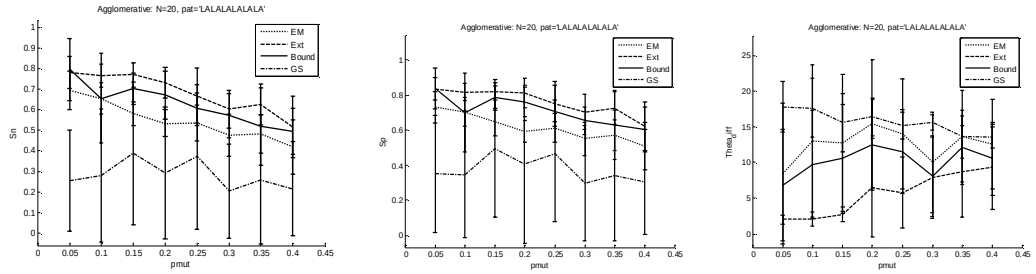
Σχήμα 5.23: Μέτρα επίδοσης για επαναληπτικό μοτίβο μήκους $K = 12$ συμβόλων μέσα σε $N = 15$ συμβολικές ακολουθίες και χρήση του συνθετικού αλγόριθμου.



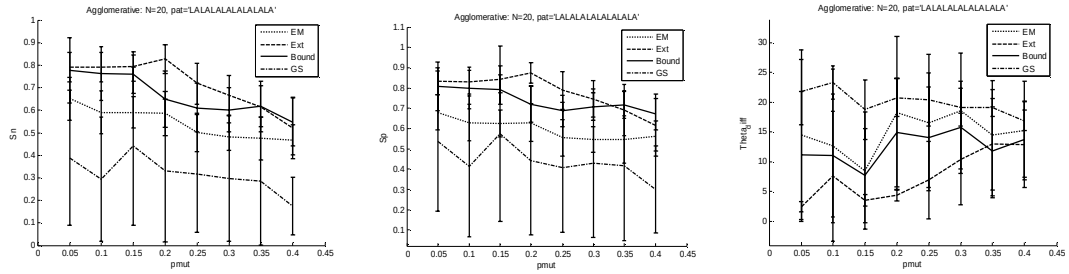
Σχήμα 5.24: Μέτρα επίδοσης για επαναληπτικό μοτίβο μήκους $K = 16$ συμβόλων μέσα σε $N = 15$ συμβολικές ακολουθίες και χρήση του συνθετικού αλγόριθμου.



Σχήμα 5.25: Μέτρα επίδοσης για επαναληπτικό μοτίβο μήκους $K = 8$ συμβόλων μέσα σε $N = 20$ συμβολικές ακολουθίες και χρήση του συνθετικού αλγόριθμου.



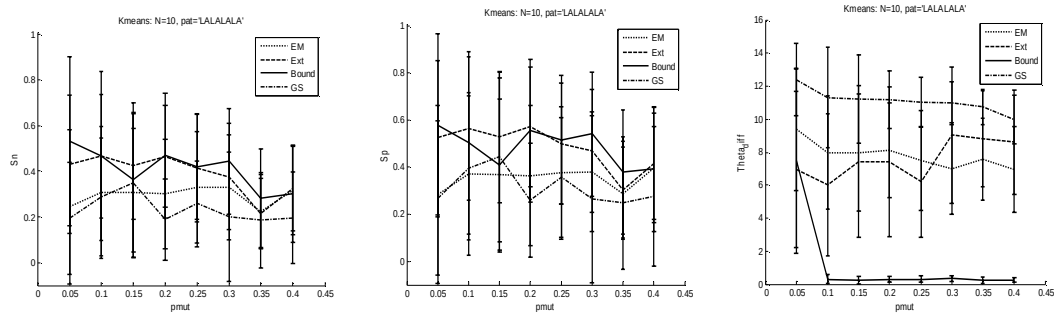
Σχήμα 5.26: Μέτρα επίδοσης για επαναληπτικό μοτίβο μήκους $K = 12$ συμβόλων μέσα σε $N = 20$ συμβολικές ακολουθίες και χρήση του συνθετικού αλγόριθμου.



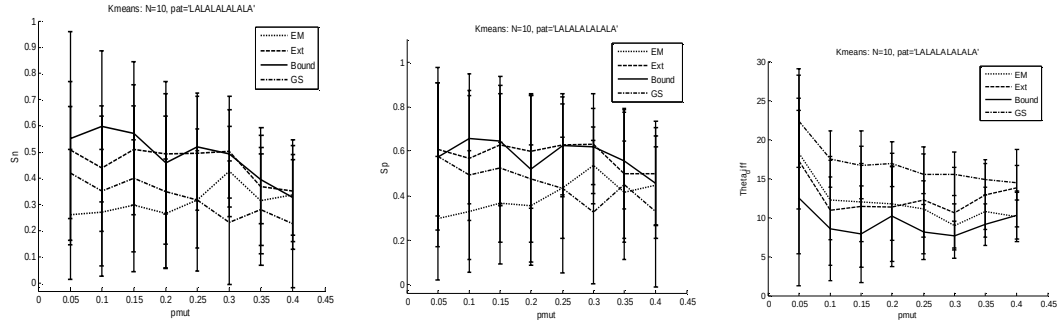
Σχήμα 5.27: Μέτρα επίδοσης για επαναληπτικό μοτίβο μήκους $K = 16$ συμβόλων μέσα σε $N = 20$ συμβολικές ακολουθίες και χρήση του συνθετικού αλγόριθμου.

Όπως συμπεραίνουμε, βασιζόμενοι στα σχήματα 5.19 – 5.27, οι δύο μέθοδοι που παρουσιάσαμε αναλυτικά στο προηγούμενο κεφάλαιο (Ext και Bound), υπερτερούν σαφώς όσον αφορά όλα τα μέτρα επίδοσης τόσο από τον αλγόριθμο EM όσο και από την μέθοδο Δειγματοληψίας του Gibbs, ενώ στη γενική περίπτωση τις καλύτερες επιδόσεις έχει και πάλι, η μέθοδος του εκτεταμένου πίνακα (Ext). Επομένως, όταν έχουμε επαναληπτικό μοτίβο, οι αλγόριθμοι EM και GS παρουσιάζουν σημαντική αδυναμία ως προς τον εντοπισμό των υπακολουθιών μήκους K, οι οποίες αποτελούν αντίγραφο του μοτίβου που αναζητούμε.

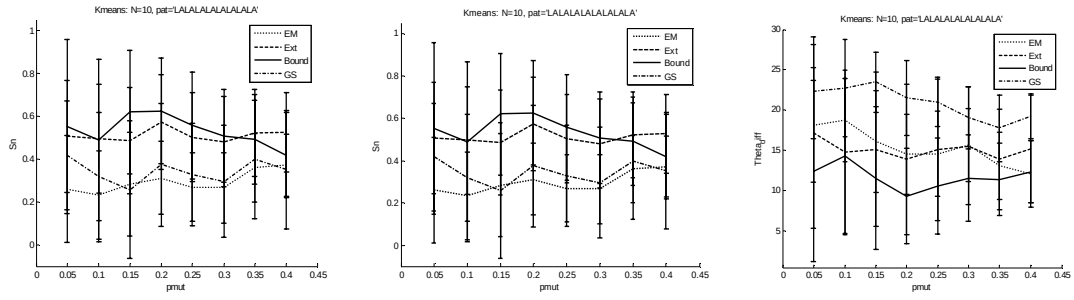
5.2.4. Επαναληπτικό Μοτίβο και Αρχικοποίηση των K-κέντρων



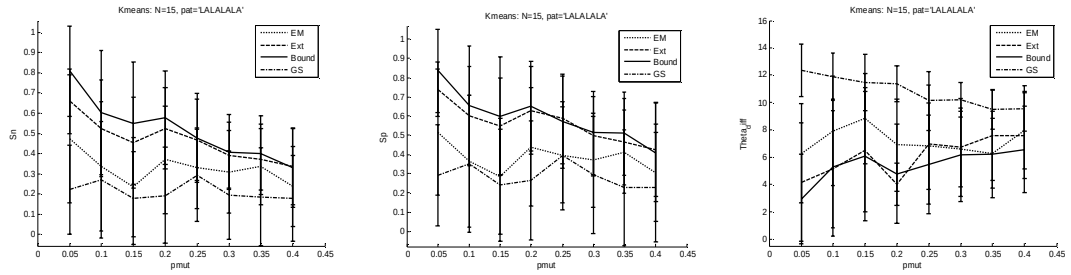
Σχήμα 5.28: Μέτρα επίδοσης για επαναληπτικό μοτίβο μήκους $K = 8$ συμβόλων μέσα σε $N = 10$ συμβολικές ακολουθίες και χρήση του αλγορίθμου των K-κέντρων.



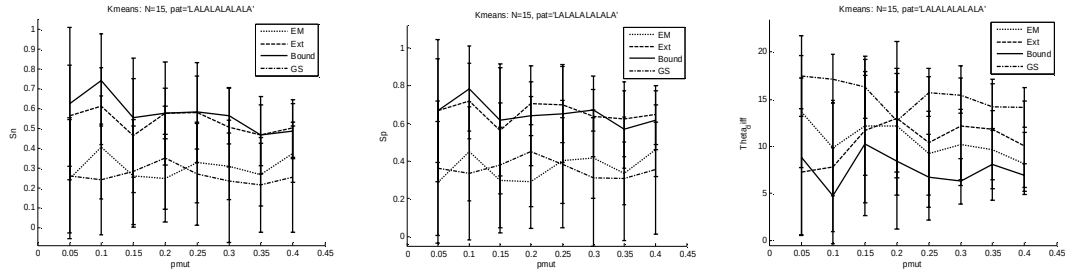
Σχήμα 5.29: Μέτρα επίδοσης για επαναληπτικό μοτίβο μήκους $K = 12$ συμβόλων μέσα σε $N = 10$ συμβολικές ακολουθίες και χρήση του αλγορίθμου των K-κέντρων.



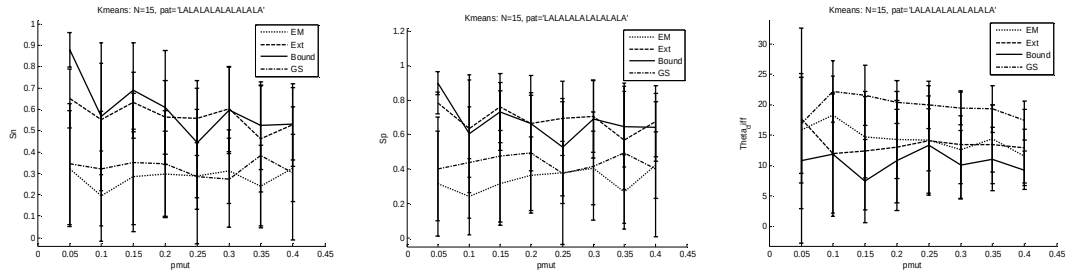
Σχήμα 5.30: Μέτρα επίδοσης για επαναληπτικό μοτίβο μήκους $K = 16$ συμβόλων μέσα σε $N = 10$ συμβολικές ακολουθίες και χρήση του αλγορίθμου των K-κέντρων.



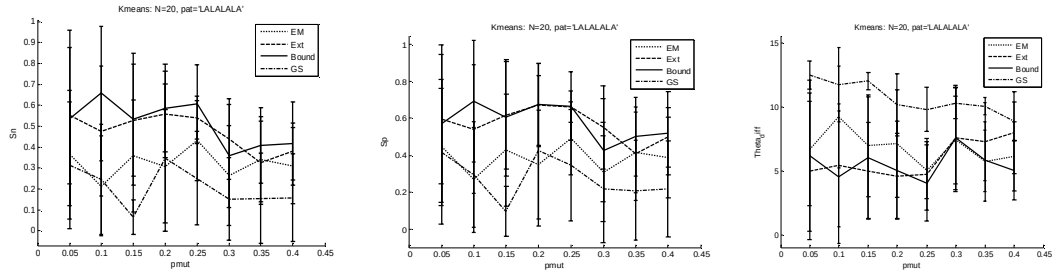
Σχήμα 5.31: Μέτρα επίδοσης για επαναληπτικό μοτίβο μήκους $K = 8$ συμβόλων μέσα σε $N = 15$ συμβολικές ακολουθίες και χρήση του αλγορίθμου των K-κέντρων.



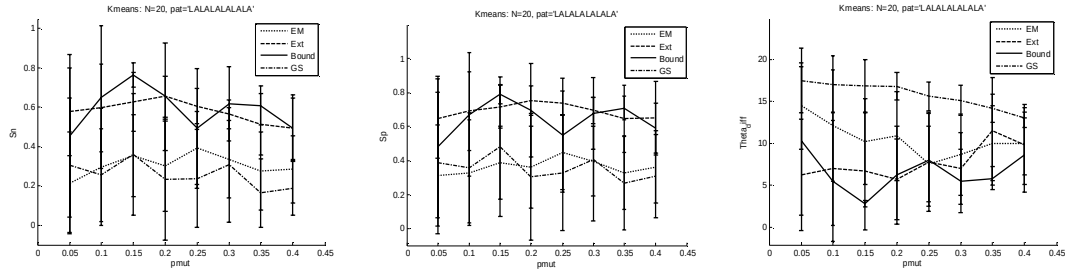
Σχήμα 5.32: Μέτρα επίδοσης για επαναληπτικό μοτίβο μήκους $K = 12$ συμβόλων μέσα σε $N = 15$ συμβολικές ακολουθίες και χρήση του αλγορίθμου των K-κέντρων.



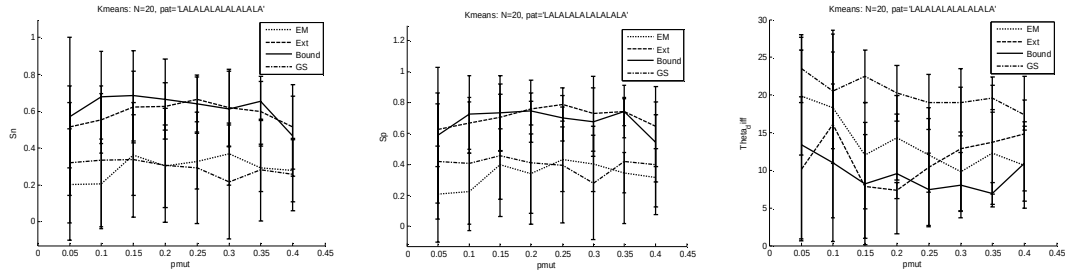
Σχήμα 5.33: Μέτρα επίδοσης για επαναληπτικό μοτίβο μήκους $K = 16$ συμβόλων μέσα σε $N = 15$ συμβολικές ακολουθίες και χρήση του αλγορίθμου των K-κέντρων.



Σχήμα 5.34: Μέτρα επίδοσης για επαναληπτικό μοτίβο μήκους $K = 8$ συμβόλων μέσα σε $N = 20$ συμβολικές ακολουθίες και χρήση του αλγορίθμου των K-κέντρων.



Σχήμα 5.35: Μέτρα επίδοσης για επαναληπτικό μοτίβο μήκους $K = 12$ συμβόλων μέσα σε $N = 20$ συμβολικές ακολουθίες και χρήση του αλγορίθμου των K-κέντρων.



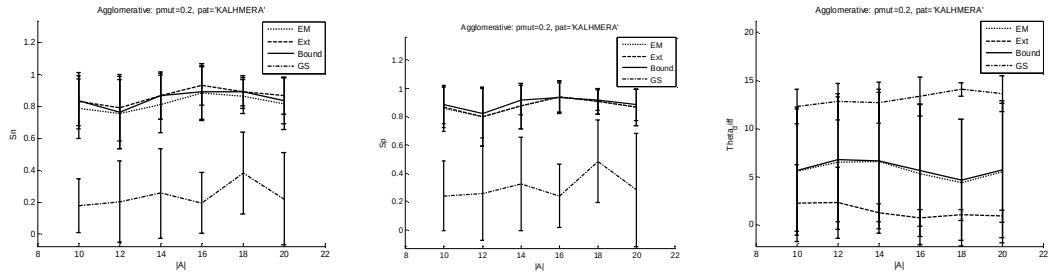
Σχήμα 5.36: Μέτρα επίδοσης για επαναληπτικό μοτίβο μήκους $K = 16$ συμβόλων μέσα σε $N = 20$ συμβολικές ακολουθίες και χρήση του αλγορίθμου των K-κέντρων.

Χρησιμοποιώντας, τώρα, την αρχικοποίηση των K-κέντρων (K-means) στην προσπάθεια προσδιορισμού του επαναληπτικού μοτίβου, και πάλι οι δύο νέες μέθοδοι του Ext (εκτεταμένου πίνακα) και Bound (οριοθετούμενου πίνακα) υπερισχύουν έναντι των EM και

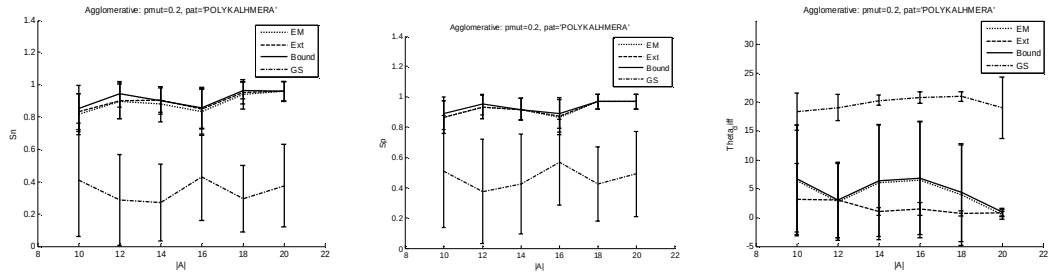
GS σε όλα τα μέτρα επίδοσης. Όμως, αυτό που αλλάζει και κρίνεται αναγκαίο να τονίσουμε είναι ότι ο αλγόριθμος του οριοθετούμενου πίνακα δίνει καλύτερα αποτελέσματα και στα ποσοστά εντοπισμού του μοτίβου Sn και Sp, αλλά πολύ περισσότερο προσφέρει καλύτερη προσέγγιση του μοτίβου (δηλαδή του παράγοντα ενδιαφέροντος θ). Επιπροσθέτως, δεν είναι λίγες οι περιπτώσεις που παρατηρούμε την αδυναμία του EM σε πιο τυχαία αρχικοποίηση από αυτή του συνθετικού αλγορίθμου, καθώς υστερεί στα μέτρα επίδοσης συγκρινόμενος ακόμα και με τον GS.

5.2.5. Πειράματα με Διαφορετικό Μέγεθος Αλφάβητου

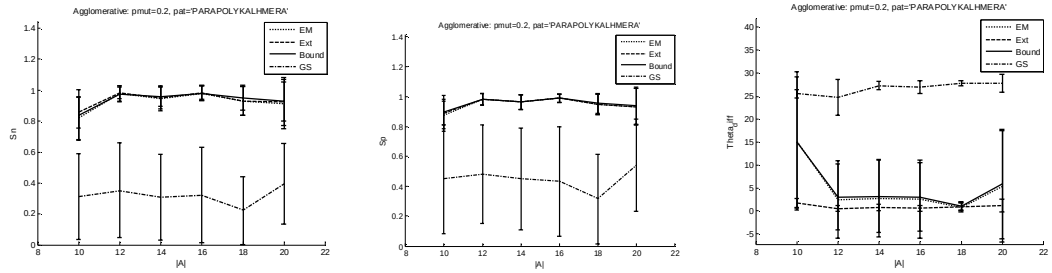
Στα γραφήματα που ακολουθούν, αναπαρίστανται τα αποτελέσματα των μέτρων επίδοσης των τεσσάρων μεθόδων για διαφορετικό μήκος του αλφάβητου ($|A|$=10, 12, 14, 16, 18, 20) και για πιθανότητα μετάλλαξης $p_{mut} = 0.2,\ 0.3$, χρησιμοποιώντας αρχικοποίηση του συνθετικού αλγόριθμου (Agglomerative), η οποία έχουμε διαπιστώσει ότι ευνοεί την επίδοση του EM.



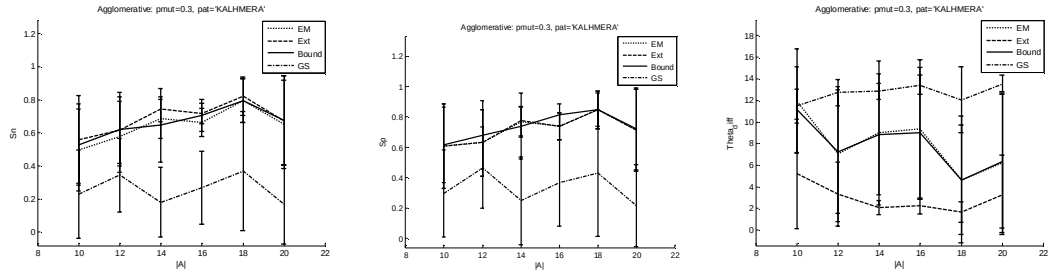
Σχήμα 5.37: Μέτρα επίδοσης για μη επαναληπτικό μοτίβο μήκους $K = 8$ συμβόλων, πιθανότητα μετάλλαξης 0.2 και χρήση του συνθετικού αλγόριθμου.



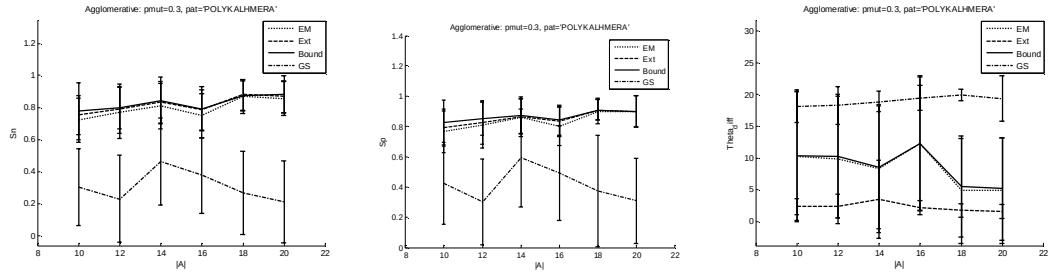
Σχήμα 5.38: Μέτρα επίδοσης για μη επαναληπτικό μοτίβο μήκους $K = 12$ συμβόλων, πιθανότητα μετάλλαξης 0.2 και χρήση του συνθετικού αλγόριθμου.



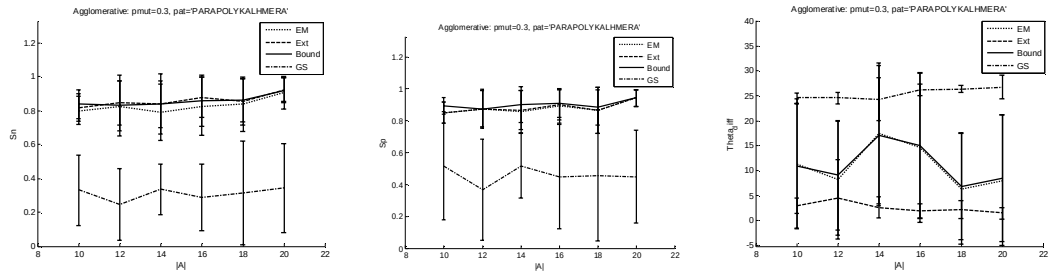
Σχήμα 5.39: Μέτρα επίδοσης για μη επαναληπτικό μοτίβο μήκους $K = 16$ συμβόλων, πιθανότητα μετάλλαξης 0.2 και χρήση του συνθετικού αλγόριθμου.



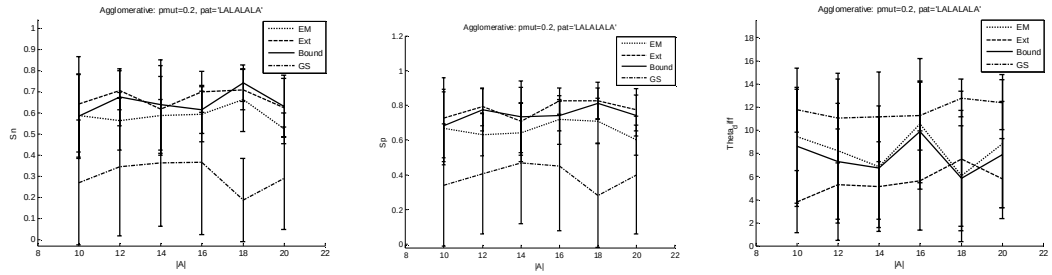
Σχήμα 5.40: Μέτρα επίδοσης για μη επαναληπτικό μοτίβο μήκους $K = 8$ συμβόλων, πιθανότητα μετάλλαξης 0.3 και χρήση του συνθετικού αλγόριθμου.



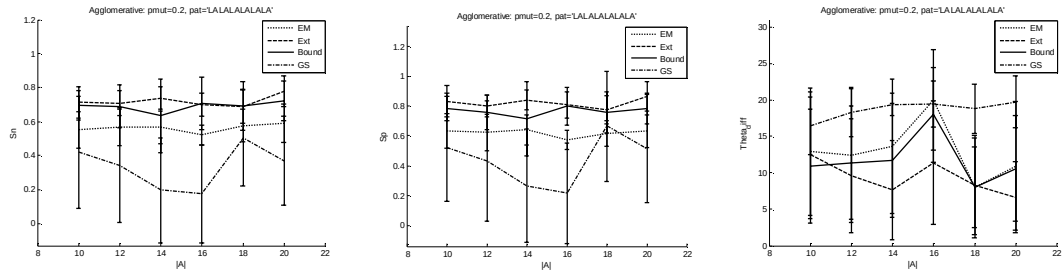
Σχήμα 5.41: Μέτρα επίδοσης για μη επαναληπτικό μοτίβο μήκους $K = 12$ συμβόλων, πιθανότητα μετάλλαξης 0.3 και χρήση του συνθετικού αλγόριθμου.



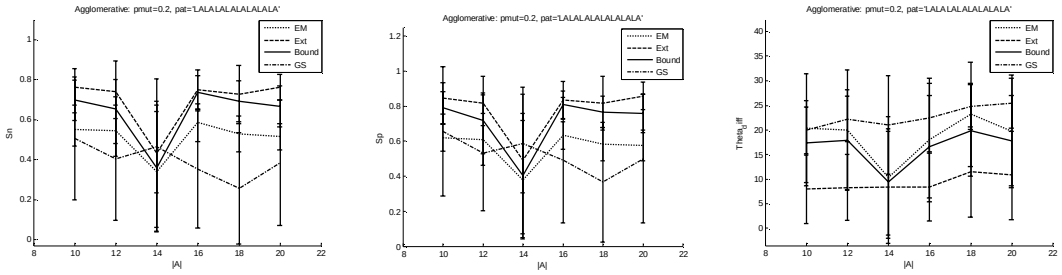
Σχήμα 5.42: Μέτρα επίδοσης για μη επαναληπτικό μοτίβο μήκους $K = 16$ συμβόλων, πιθανότητα μετάλλαξης 0.3 και χρήση του συνθετικού αλγόριθμου.



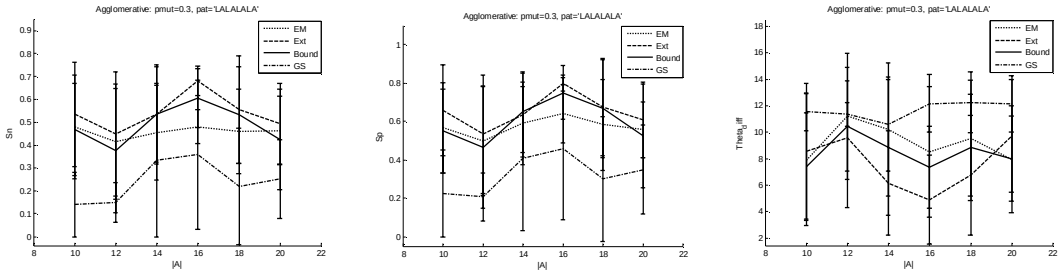
Σχήμα 5.43: Μέτρα επίδοσης για επαναληπτικό μοτίβο μήκους $K = 8$ συμβόλων, πιθανότητα μετάλλαξης 0.2 και χρήση του συνθετικού αλγόριθμου.



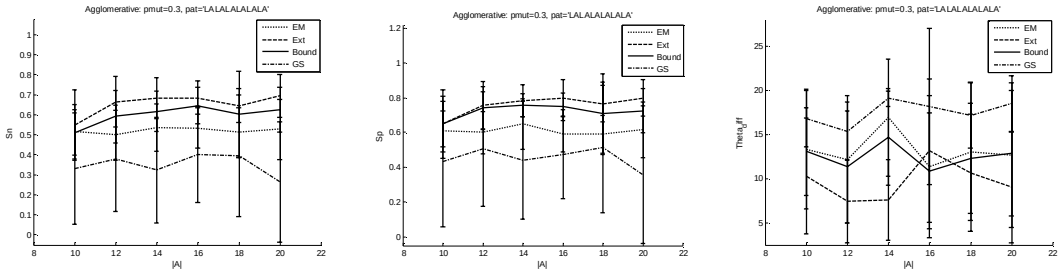
Σχήμα 5.44: Μέτρα επίδοσης για επαναληπτικό μοτίβο μήκους $K = 12$ συμβόλων, πιθανότητα μετάλλαξης 0.2 και χρήση του συνθετικού αλγόριθμου.



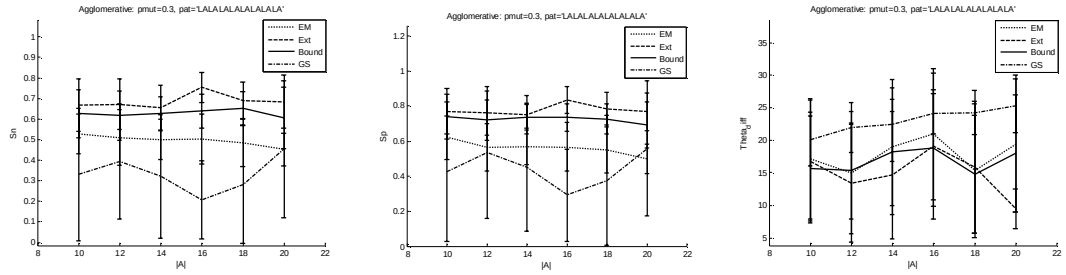
Σχήμα 5.45: Μέτρα επίδοσης για επαναληπτικό μοτίβο μήκους $K = 16$ συμβόλων, πιθανότητα μετάλλαξης 0.2 και χρήση του συνθετικού αλγόριθμου.



Σχήμα 5.46: Μέτρα επίδοσης για επαναληπτικό μοτίβο μήκους $K = 8$ συμβόλων, πιθανότητα μετάλλαξης 0.3 και χρήση του συνθετικού αλγόριθμου.



Σχήμα 5.47: Μέτρα επίδοσης για επαναληπτικό μοτίβο μήκους $K = 12$ συμβόλων, πιθανότητα μετάλλαξης 0.3 και χρήση του συνθετικού αλγόριθμου.



Σχήμα 5.48: Μέτρα επίδοσης για επαναληπτικό μοτίβο μήκους $K = 16$ συμβόλων, πιθανότητα μετάλλαξης 0.3 και χρήση του συνθετικού αλγόριθμου.

Όπως διαπιστώνουμε, παρατηρώντας τα σχήματα 5.37 – 5.42, στην περίπτωση του μη επαναληπτικού μοτίβου, οι δύο νέες μέθοδοι υπερισχύουν για άλλη μια φορά έναντι του EM στα ποσοστά Sn και Sp, έστω και με μικρή διαφορά. Παρά το γεγονός αυτό, ο EM δίνει συνήθως καλύτερη προσέγγιση του παράγοντα ενδιαφέροντος (δηλαδή μικρότερη μέση τιμή του μέτρου $\Delta\theta$) από τον αλγόριθμο του οριοθετούμενου πίνακα. Και πάλι, όμως, η μέθοδος του εκτεταμένου πίνακα δίνει τις καλύτερες τιμές στα μέτρα επίδοσης.

Οι παραπάνω παρατηρήσεις ανατρέπονται στην περίπτωση του επαναληπτικού μοτίβου, όπου πια οι μέθοδοι Ext (εκτεταμένου πίνακα) και Bound (οριοθετούμενου πίνακα) έχουν μια σαφή διαφορά από τον EM, ο οποίος δείχνει αδύναμος να το εντοπίσει. Τα αποτελέσματα δεν παρουσιάζουν κάποια σημαντική αλλαγή για τις διαφορετικές τιμές (0.2 και 0.3) της πιθανότητας μετάλλαξης, ενώ είναι σημαντικό να σημειώσουμε ότι σε κάποιες περιπτώσεις η μέθοδος του οριοθετούμενου πίνακα δίνει καλύτερη τιμή στο μέτρο $\Delta\theta$.

Τέλος, ο αλγόριθμος Δειγματοληψίας του Gibbs έχει, και πάλι, στη γενική περίπτωση τα χαμηλότερα μέτρα επίδοσης όσον αφορά τον εντοπισμό του μοτίβου. Παρ' όλα αυτά, για επαναληπτικό μοτίβο και μεγάλες τιμές στο μήκος του αλφάβητου, υπάρχουν και περιπτώσεις όπου υπερτερεί του EM στα μέτρα που εξετάζουμε.

5.3. Πειραματική Μελέτη σε Πραγματικά Δεδομένα

Προκειμένου να εξάγουμε πιο σαφή και ολοκληρωμένα συμπεράσματα πάνω στην επίδοση των αλγορίθμων EM, Ext, Bound και GS, διεξήχθησαν πειράματα χρησιμοποιώντας

πραγματικά δεδομένα, τα οποία προέρχονται από το site http://dragon.bio.purdue.edu/pmotif. Συγκεκριμένα, πρόκειται για δεδομένα που βρίσκονται στο εργαστήριο Βιοπληροφορικής του Τμήματος Πληροφορικής του Πανεπιστημίου Purdue και χρησιμοποιούνται συχνά για την αξιολόγηση αλγορίθμων ανίχνευσης μοτίβων, καθώς είναι γνωστό πού ακριβώς βρίσκεται το μοτίβο σε κάθε συμβολοσειρά (ground truth). Μάλιστα, αξίζει να σημειώσουμε ότι ο βαθμός ομοιότητας μεταξύ των μοτίβων (σε κάθε σύνολο συμβολοσειρών που διατίθεται) είναι πολύ χαμηλός, γεγονός που ενισχύει το ενδιαφέρον για τα αποτελέσματα των πειραμάτων που θα παρουσιάσουμε στη συνέχεια.

Εδώ, λοιπόν, διακρίνουμε δύο κατηγορίες συμβολοσειρών, οι οποίες αποτελούνται από σύμβολα του αλφάβητου $A = [\text{AGCT}]$ και συνεπώς, συνιστούν αλυσίδες DNA (δηλαδή πρόκειται για βιολογικά δεδομένα). Στην πρώτη κατηγορία ανήκουν συμβολοσειρές στις οποίες το μοτίβο εκτείνεται από την 21^{η} θέση και τελειώνει 20 θέσεις πριν από το τελευταίο τους σύμβολο. Στη δεύτερη, τώρα, κατηγορία ανήκουν μεγαλύτερου μήκους αλυσίδες, όπου το μοτίβο αρχίζει από την 51^{η} θέση τους και τελειώνει 50 θέσεις πριν το τελευταίο σύμβολο.

Στους πίνακες που ακολουθούν παρουσιάζονται, για κάθε σύνολο συμβολοσειρών, το μήκος του εκάστοτε μοτίβου, η μέση απόδοσή του στα μέτρα Sn, Sp και $\Delta\theta$, καθώς και η τυπική τους απόκλιση με βάση το σύνολο των πειραμάτων που πραγματοποιήθηκαν. Σημειωτέον ότι, στην προκειμένη περίπτωση, χρησιμοποιήθηκε τυχαία αρχικοποίηση για τον πίνακα περιγραφής του μοτίβου θ και αρχικοποίηση με τον αλγόριθμο των K-κέντρων. Με αυτό τον τρόπο, επιθυμούμε να διαπιστώσουμε κατά πόσο είναι αποτελεσματικές οι τέσσερις μέθοδοι που χρησιμοποιήσαμε, χωρίς και υπό την «εύνοια» κάποιου αλγόριθμου αρχικοποίησης. Μια ανάλογη και εξαιρετικά ενδιαφέρουσα μελέτη αξιολόγησης κάποιων αλγορίθμων σε πραγματικά βιολογικά σύνολα δεδομένων περιγράφεται στην [11].

Πίνακας 5.1: Αποτελέσματα για την 1η Κατηγορία Αλυσίδων DNA με Τυχαία Αρχικοποίηση.

Σύνολο αλυσίδων	Μήκος Μοτίβου	Μέτρο επίδοσης	Μέθοδος			
			Ext	*Bound*	*EM*	*GS*
AraC20	17	*Sn*	0.2892 ± 0.0902	0.3124 ± 0.0559	0.1898 ± 0.1308	**0.4722 ± 0.2216**
		Sp	0.4021 ± 0.1227	0.4521 ± 0.0900	0.2673 ± 0.1772	**0.5931 ± 0.3180**
		$\Delta\theta$	16.6469 ± 2.5852	11.4239 ± 3.0945	10.7538 ± 4.3391	**9.9983 ± 2.3570**
ArcA20	61	*Sn*	0.1020 ± 0.0600	0.5957 ± 0.1162	0.2246 ± 0.1522	**0.8510 ± 0.0778**
		Sp	0.1208 ± 0.0689	0.7062 ± 0.1202	0.2750 ± 0.1831	**1.0000 ± 0.0000**
		$\Delta\theta$	54.0200 ± 6.0300	38.3741 ± 12.8260	25.8774 ± 11.5069	**19.7517 ± 1.3091**
ArgR20	16	*Sn*	0.4049 ± 0.1148	0.3749 ± 0.0695	0.3115 ± 0.1321	**0.4744 ± 0.2296**
		Sp	0.5555 ± 0.1142	0.5486 ± 0.1093	0.4306 ± 0.1771	**0.6736 ± 0.3472**
		$\Delta\theta$	16.3497 ± 1.5068	12.9520 ± 2.4392	12.4445 ± 3.3679	**10.6728 ± 2.0653**
CytR20	40	*Sn*	0.2479 ± 0.1398	0.5388 ± 0.0967	0.3676 ± 0.0675	**0.7767 ± 0.0817**
		Sp	0.3417 ± 0.1782	0.7167 ± 0.1337	0.5250 ± 0.0965	**1.0000 ± 0.0000**
		$\Delta\theta$	40.3325 ± 7.2843	33.4239 ± 8.5023	29.3123 ± 7.4817	**24.7386 ± 1.9502**

DnaA20	9	*Sn*	**0.3267 ± 0.1094**	0.2526 ± 0.1043	0.3042 ± 0.0822	0.2354 ± 0.1941
		Sp	**0.5357 ± 0.1379**	0.3929 ± 0.2121	0.4881 ± 0.1771	0.3095 ± 0.2424
		$\Delta\theta$	8.6290 ± **0.6367**	7.8070 ± 1.5223	8.0436 ± 1.2816	**7.4707** ± 1.6799
Fur20	19	*Sn*	0.4144 ± 0.1475	0.3326 ± 0.0472	0.2897 ± 0.0625	**0.3645 ± 0.2386**
		Sp	0.6000 ± 0.2078	0.5125 ± 0.0608	0.4000 ± 0.0929	**0.4708 ± 0.3461**
		$\Delta\theta$	18.0587 ± 1.7872	13.8258 ± 3.6983	13.0923 ± 4.3131	**10.8241 ± 2.1148**
FruR20	14	*Sn*	0.8576 ± 0.1957	0.6153 ± 0.2290	0.6189 ± 0.3211	**0.4845 ± 0.2588**
		Sp	0.9083 ± 0.1730	0.7583 ± 0.2275	0.7000 ± 0.3219	**0.7667 ± 0.3339**
		$\Delta\theta$	11.1317 ± 4.8618	15.9920 ± 3.2859	14.5381 ± 3.7681	**14.4436 ± 1.4975**
FNR20	22	*Sn*	0.5271 ± 0.0814	0.5504 ± 0.1393	0.1142 ± 0.1312	**0.5127 ± 0.1966**
		Sp	0.6387 ± 0.0877	0.7440 ± 0.1486	0.1467 ± 0.1545	**0.7720 ± 0.3667**
		$\Delta\theta$	16.2620 ± 2.8337	14.3055 ± 1.3665	11.6925 ± 2.9859	**10.6842 ± 0.7851**
GlpR20	20	*Sn*	0.3969 ± 0.1519	0.4686 ± 0.0650	0.2507 ± 0.1069	**0.4790 ± 0.2165**
		Sp	0.4956 ± 0.1716	0.5965 ± 0.0518	0.3290 ± 0.1311	**0.7105 ± 0.4036**
		$\Delta\theta$	17.7458 ± 5.0811	17.8199 ± 4.2062	16.1666 ± 3.2505	**15.2068 ± 1.8358**

LexA20	20	*Sn*	0.5498 ± 0.1514	0.4626 ± 0.1276	0.3128 ± 0.1361	**0.5650 ± 0.2483**
		Sp	0.6583 ± 0.1929	0.6833 ± 0.1115	0.4083 ± 0.1564	**0.7667 ± 0.3257**
		$\Delta\theta$	23.1515 ± 1.6738	17.7087 ± 1.8744	16.5336 ± 3.7040	**12.8980 ± 2.2707**
Nac20	15	*Sn*	0.2543 ± 0.0795	0.2835 ± 0.0416	0.0573 ± 0.0804	**0.3895 ± 0.1936**
		Sp	0.3426 ± 0.1204	0.4166 ± 0.0502	0.1018 ± 0.1294	**0.5000 ± 0.2940**
		$\Delta\theta$	14.9149 ± 2.1080	11.3682 ± 2.0144	9.5207 ± 3.1099	**9.9827 ± 1.1546**
NtrC20	15	*Sn*	0.4396 ± 0.0683	0.3113 ± 0.0969	0.2632 ± 0.1393	**0.4006 ± 0.2228**
		Sp	0.5649 ± 0.0321	0.4908 ± 0.1107	0.3704 ± 0.1664	**0.6667 ± 0.3790**
		$\Delta\theta$	7.8368 ± 4.3257	16.1063 ± 3.4910	15.7821 ± 5.1392	**13.6949 ± 3.1241**
OmpR20	10	*Sn*	0.2421 ± 0.1100	0.1576 ± 0.0919	0.1367 ± 0.0804	**0.3050 ± 0.3030**
		Sp	0.3611 ± 0.1669	0.2500 ± 0.1823	0.2000 ± 0.1531	**0.4389 ± 0.3902**
		$\Delta\theta$	8.2852 ± 1.2274	6.4986 ± 1.4659	7.6352 ± 1.4267	**5.8682 ± 0.8441**
OxyR20	45	*Sn*	0.2221 ± 0.1187	0.5676 ± 0.1449	0.3165 ± 0.1640	**0.7340 ± 0.0696**
		Sp	0.3055 ± 0.1508	0.7222 ± 0.1535	0.4167 ± 0.2017	**1.0000 ± 0.0000**
		$\Delta\theta$	47.4697 ± 8.6334	39.0084 ± 5.9596	28.0299 ± 6.6362	**29.5398 ± 1.0166**

PhoB20	17	*Sn*	0.3365 ± 0.1618	0.3426 ± 0.0869	0.2060 ± 0.1469	**0.4761 ± 0.2641**
		Sp	0.4611 ± 0.2155	0.5000 ± 0.1189	0.2722 ± 0.2155	**0.6778 ± 0.3301**
		$\Delta\theta$	15.5442 ± 2.7740	14.2403 ± 2.0565	12.2610 ± 2.7620	**10.9200 ± 2.9771**
PurR20	16	*Sn*	0.8048 ± 0.0351	0.4987 ± 0.2162	0.2554 ± 0.2526	**0.4576 ± 0.3037**
		Sp	0.8778 ± 0.0385	0.6667 ± 0.1658	0.3611 ± 0.2659	**0.6444 ± 0.3880**
		$\Delta\theta$	11.9603 ± 2.8923	14.5164 ± 1.5931	12.4479 ± 2.1120	**12.2689 ± 2.8151**
PurR20	16	*Sn*	0.8048 ± 0.0351	0.4987 ± 0.2162	0.2554 ± 0.2526	**0.4576 ± 0.3037**
		Sp	0.8778 ± 0.0385	0.6667 ± 0.1658	0.3611 ± 0.2659	**0.6444 ± 0.3880**
		$\Delta\theta$	11.9603 ± 2.8923	14.5164 ± 1.5931	12.4479 ± 2.1120	**12.2689 ± 2.8151**
SoxS20	18	*Sn*	0.3506 ± 0.0974	0.3385 ± 0.0698	0.1327 ± 0.0739	**0.4417 ± 0.2173**
		Sp	0.4755 ± 0.1407	0.4951 ± 0.1048	0.1912 ± 0.0943	**0.6765 ± 0.3342**
		$\Delta\theta$	15.8202 ± 2.1419	11.5134 ± 1.6624	10.6930 ± 3.5480	**9.2307 ± 1.1990**
TyrR20	22	*Sn*	0.4254 ± 0.1441	0.4666 ± 0.1979	0.2264 ± 0.1523	**0.5134 ± 0.2291**
		Sp	0.5588 ± 0.1515	0.6078 ± 0.1762	0.3333 ± 0.2028	**0.6863 ± 0.3815**
		$\Delta\theta$	21.2849 ± 3.2647	16.4993 ± 3.6364	15.6823 ± 3.6663	**14.0014 ± 3.0210**

Πίνακας 5.2: Αποτελέσματα για την 1η Κατηγορία Αλυσίδων DNA με Αρχικοποίηση Κ-κέντρων.

Σύνολο αλυσίδων	Μήκος Μοτίβου	Μέτρο επίδοσης	Μέθοδος			
			Ext	*Bound*	*EM*	*GS*
AraC20	17	*Sn*	0.3309 ± 0.0706	0.2884 ± 0.1229	0.3669 ± 0.0694	**0.4300 ± 0.2001**
		Sp	0.4681 ± 0.0818	0.4292 ± 0.1737	0.5555 ± 0.0894	**0.5910 ± 0.3254**
		$\Delta\theta$	12.9110 ± 0.9845	16.8949 ± 2.0823	11.5501 ± 1.4555	**10.9990 ± 2.2930**
ArcA20	61	*Sn*	0.2115 ± 0.1195	0.0812 ± 0.0588	0.5790 ± 0.0784	**0.8414 ± 0.0919**
		Sp	0.2562 ± 0.1386	0.0979 ± 0.0695	0.6896 ± 0.0882	**1.0000 ± 0.0000**
		$\Delta\theta$	19.1072 ± 4.5075	58.0077 ± 9.2013	30.3413 ± 4.2130	**20.5066 ± 1.2591**
ArgR20	16	*Sn*	0.2301 ± 0.1234	0.3390 ± 0.1337	0.3216 ± 0.0767	**0.3932 ± 0.2102**
		Sp	0.3542 ± 0.1816	0.4375 ± 0.1709	0.5278 ± 0.1521	**0.5903 ± 0.3046**
		$\Delta\theta$	12.8059 ± 2.4245	15.0260 ± 2.9189	11.0507 ± 3.2654	**11.1245 ± 1.2007**
CytR20	40	*Sn*	0.4717 ± 0.0811	0.3303 ± 0.1603	0.5566 ± 0.0991	**0.6965 ± 0.1141**

		Sp	0.6333 ± 0.1073	0.4583 ± 0.2193	0.7250 ± 0.1138	**1.0000 ± 0.0000**
		$\Delta\theta$	29.0604 ± 2.9003	39.8575 ± 4.6601	32.0081 ± 3.7108	**24.6847 ± 2.2706**
DnaA20	9	*Sn*	0.1903 ± 0.1657	0.2495 ± 0.1793	0.2427 ± 0.1178	**0.2818 ± 0.2561**
		Sp	0.2976 ± 0.2316	0.3809 ± 0.2462	0.4048 ± 0.2178	**0.4167 ± 0.3073**
		$\Delta\theta$	9.4136 ± 1.5890	9.4587 ± 1.6160	8.0908 ± 1.4241	**8.1837 ± 0.3073**
Fur20	19	*Sn*	0.3223 ± 0.0974	**0.4198 ± 0.1464**	0.3457 ± 0.0602	0.4081 ± 0.2424
		Sp	0.4792 ± 0.1453	**0.5875 ± 0.1860**	0.5542 ± 0.0811	0.5500 ± 0.3784
		$\Delta\theta$	14.9146 ± 2.7572	18.3114 ± 2.7349	11.4069 ± 3.1583	**11.3135 ± 1.6577**
FruR20	14	*Sn*	0.3868 ± 0.3776	0.5527 ± 0.3089	**0.7316 ± 0.2588**	0.4756 ± 0.2815
		Sp	0.4417 ± 0.3919	0.6333 ± 0.3200	**0.8333 ± 0.2146**	0.7083 ± 0.3343
		$\Delta\theta$	11.6069 ± 3.7599	12.8390 ± 4.3484	**8.8420 ± 3.9023**	13.3696 ± 4.0078
FNR20	22	*Sn*	0.2766 ± 0.1688	**0.5710 ± 0.0407**	0.5391 ± 0.1718	0.4458 ± 0.2241
		Sp	0.3560 ± 0.2140	**0.7040 ± 0.0567**	0.7067 ± 0.1989	0.6453 ± 0.4181
		$\Delta\theta$	12.7685 ± 3.5835	15.5267 ± 3.1954	13.6163 ± 3.3553	**10.7063 ± 0.7812**
GlpR20	20	*Sn*	0.3440 ± 0.0766	0.3726 ± 0.1646	0.4825 ± 0.0637	**0.4853 ± 0.2325**

		Sp	0.4649 ± 0.1246	0.4649 ± 0.1941	0.6096 ± 0.0352	**0.6974 ± 0.4217**
		$\Delta\theta$	16.5459 ± 3.9727	20.1486 ± 2.9614	**14.4463 ± 5.3889**	16.0597 ± 2.2864
LexA20	20	*Sn*	0.3323 ± 0.1915	0.4332 ± 0.1909	**0.5304 ± 0.1705**	0.5004 ± 0.2202
		Sp	0.4333 ± 0.2462	0.4917 ± 0.2314	0.6583 ± 0.2151	**0.6833 ± 0.3040**
		$\Delta\theta$	17.8154 ± 2.0987	22.4104 ± 1.8167	19.3596 ± 2.3598	**13.4860 ± 1.4310**
Nac20	15	*Sn*	0.1983 ± 0.0628	0.2368 ± 0.0901	0.3163 ± 0.0729	**0.4753 ± 0.2116**
		Sp	0.2963 ± 0.1368	0.3426 ± 0.1204	0.4444 ± 0.1161	**0.6574 ± 0.3159**
		$\Delta\theta$	12.5607 ± 1.2533	14.8753 ± 1.7240	11.1465 ± 1.8115	**10.2066 ± 0.7323**
NrtC20	15	*Sn*	0.1233 ± 0.0856	0.2324 ± 0.1553	0.3542 ± 0.1196	**0.5222 ± 0.2535**
		Sp	0.1759 ± 0.1204	0.2870 ± 0.1863	0.5000 ± 0.1535	**0.6944 ± 0.3421**
		$\Delta\theta$	16.4488 ± 2.4717	15.4732 ± 5.9153	16.1059 ± 3.4800	**14.4286 ± 3.1770**
OmpR20	10	*Sn*	0.1561 ± 0.0641	0.2184 ± 0.0721	0.1503 ± 0.0500	**0.2883 ± 0.2263**
		Sp	0.2389 ± 0.1188	0.3500 ± 0.1425	0.2556 ± 0.1131	**0.4111 ± 0.3301**
		$\Delta\theta$	8.8959 ± 1.2468	9.3962 ± 1.1447	6.8211 ± 1.0917	**5.9268 ± 0.6425**
OxyR20	45	*Sn*	0.4523 ± 0.1132	0.2586 ± 0.1741	0.6433 ± 0.1167	**0.8124 ± 0.0898**

		Sp	0.5833 ± 0.1431	0.3426 ± 0.1922	0.7963 ± 0.1238	**1.0000 ± 0.0000**
		$\Delta\theta$	30.3185 ± 1.5706	43.2452 ± 5.2133	33.5580 ± 3.1041	**28.1587 ± 5.3597**
PhoB20	17	*Sn*	0.1926 ± 0.1090	0.2669 ± 0.1206	0.3194 ± 0.0682	**0.5814 ± 0.2498**
		Sp	0.2667 ± 0.1449	0.3722 ± 0.1786	0.4778 ± 0.0743	**0.7056 ± 0.3309**
		$\Delta\theta$	15.5975 ± 2.3894	15.8530 ± 2.0134	13.9960 ± 1.7211	**10.5821 ± 3.1202**
PurR20	16	*Sn*	0.3775 ± 0.2694	**0.6124 ± 0.2211**	0.5555 ± 0.1895	0.3663 ± 0.2272
		Sp	0.4889 ± 0.2931	0.7056 ± 0.2300	**0.7167 ± 0.1755**	0.6833 ± 0.3740
		$\Delta\theta$	13.5816 ± 2.3663	15.3765 ± 2.7193	13.6618 ± 2.9535	**12.6450 ± 1.3980**
SoxS20	18	*Sn*	0.2950 ± 0.0751	0.3217 ± 0.0725	0.3378 ± 0.0739	**0.5087 ± 0.3028**
		Sp	0.3971 ± 0.1153	0.4510 ± 0.0845	0.4902 ± 0.1287	**0.6373 ± 0.3761**
		$\Delta\theta$	13.4171 ± 1.5319	15.4226 ± 2.1466	11.1597 ± 1.0945	**8.7655 ± 2.2317**
TyrR20	22	*Sn*	0.3550 ± 0.1203	0.3566 ± 0.1311	0.4735 ± 0.1627	**0.6357 ± 0.2134**
		Sp	0.4951 ± 0.1240	0.4902 ± 0.1311	0.6961 ± 0.1370	**0.8873 ± 0.2135**
		$\Delta\theta$	17.0895 ± 2.3541	20.9908 ± 2.4432	17.2016 ± 2.6656	**14.2398 ± 2.7311**

Πίνακας 5.3: Αποτελέσματα για την 2η Κατηγορία Αλυσίδων DNA με Τυχαία Αρχικοποίηση.

Σύνολο αλυσίδων	Μήκος Μοτίβου	Μέτρο επίδοσης	Μέθοδος			
			Ext	*Bound*	*EM*	*GS*
AraC50	17	*Sn*	0.1148 ± 0.0987	0.1322 ± 0.1046	0.2565 ± 0.2140	**0.4617 ± 0.3986**
		Sp	0.1440 ± 0.1238	0.1477 ± 0.1359	0.3220 ± 0.2403	**0.5189 ± 0.4549**
		$\Delta\theta$	27.6916 ± 26.9536	51.0538 ± 39.2306	39.1676 ± 31.5482	**27.0355 ± 17.2309**
CytR50	40	*Sn*	0.0847 ± 0.0582	0.2715 ± 0.1088	0.3730 ± 0.1143	**0.4144 ± 0.2084**
		Sp	0.1000 ± 0.0953	0.4083 ± 0.1621	0.5917 ± 0.1929	**0.6417 ± 0.3118**
		$\Delta\theta$	27.0464 ± 3.5358	33.6973 ± 6.1492	27.7639 ± 3.9076	**26.1185 ± 1.5046**
DnaA50	9	*Sn*	**0.2094 ± 0.2020**	0.2909 ± 0.1594	0.0780 ± 0.0994	0.1056 ± 0.1865
		Sp	**0.3399 ± 0.2738**	0.3941 ± 0.1934	0.0985 ± 0.1214	0.1478 ± 0.2792
		$\Delta\theta$	8.1368 ± 1.2738	7.9998 ± 1.9169	**7.4125 ± 0.9340**	7.6564 ± 1.3663
LexA50	20	*Sn*	0.2576 ± 0.2079	**0.5139 ± 0.1811**	0.2925 ± 0.1548	0.1921 ± 0.2110
		Sp	0.3118 ± 0.2571	**0.6176 ± 0.2481**	0.3706 ± 0.1759	0.3118 ± 0.3740

		$\Delta\theta$	15.9865 ± 3.2297	20.1537 ± 3.0055	15.1043 ± 2.1404	**14.7814 ± 1.2290**
Nac50	15	*Sn*	0.0656 ± 0.1387	0.1481 ± 0.0773	0.1800 ± 0.1866	**0.3028 ± 0.3263**
		Sp	0.0606 ± 0.1670	0.1868 ± 0.0993	0.2172 ± 0.2182	**0.3687 ± 0.4019**
		$\Delta\theta$	**13.8053 ± 13.4917**	26.6745 ± 30.0535	19.2016 ± 17.1291	17.4820 ± 15.3945
Fur50	19	*Sn*	0.1153 ± 0.1249	**0.3671 ± 0.1201**	0.3053 ± 0.1340	0.2351 ± 0.2896
		Sp	0.1500 ± 0.1648	**0.5133 ± 0.1727**	0.4400 ± 0.1442	0.2700 ± 0.3595
		$\Delta\theta$	16.1334 ± 12.0157	26.9795 ± 30.3080	19.4887 ± 17.8717	**15.3838 ± 11.0188**
FruR50	14	*Sn*	0.2224 ± 0.3408	**0.8003 ± 0.1091**	0.1614 ± 0.1499	0.1442 ± 0.2407
		Sp	0.2364 ± 0.3828	**0.8909 ± 0.0831**	0.2273 ± 0.2370	0.2000 ± 0.3194
		$\Delta\theta$	13.8736 ± 2.7674	**12.2470 ± 5.5970**	13.2744 ± 1.3826	14.0629 ± 1.5478
GlpR50	80	*Sn*	0.2214 ± 0.1502	0.1608 ± 0.0472	0.4748 ± 0.2052	**0.8643 ± 0.0552**
		Sp	0.2594 ± 0.1720	0.1842 ± 0.0536	0.5488 ± 0.2387	**1.0000 ± 0.0000**
		$\Delta\theta$	60.8758 ± 27.5870	86.1320 ± 13.7407	66.1866 ± 18.6494	**45.4661 ± 2.2963**
NtrC50	15	*Sn*	0.2081 ± 0.0560	0.2306 ± 0.0971	**0.2307 ± 0.1077**	0.1158 ± 0.1038

		Sp	0.2727 ± 0.0764	**0.3030 ± 0.1121**	0.2929 ± 0.1141	0.1818 ± 0.2345
		$\Delta\theta$	13.4470 ± 4.9605	**13.0473 ± 6.5000**	14.9812 ± 4.0089	13.9317 ± 2.5144
OmpR50	10	*Sn*	0.0481 ± 0.0304	0.1311 ± 0.0428	0.0989 ± 0.0513	**0.1546 ± 0.1759**
		Sp	0.0909 ± 0.0616	0.2364 ± 0.1005	0.1697 ± 0.1090	**0.2242 ± 0.2276**
		$\Delta\theta$	7.0659 ± 0.8527	8.5389 ± 1.7084	6.3746 ± 1.2756	**5.8757 ± 1.2996**
OxyR50	45	*Sn*	0.1691 ± 0.0675	0.1991 ± 0.1470	0.3386 ± 0.0706	**0.4355 ± 0.1103**
		Sp	0.2020 ± 0.0670	0.3030 ± 0.2336	0.4849 ± 0.1027	**0.6263 ± 0.2345**
		$\Delta\theta$	25.1863 ± 7.1920	40.3716 ± 8.6587	32.8305 ± 5.6841	**30.3400 ± 1.8645**
PhoB50	17	*Sn*	0.0963 ± 0.1218	**0.2501 ± 0.0615**	0.1903 ± 0.0769	0.1654 ± 0.2263
		Sp	0.1212 ± 0.1424	**0.3576 ± 0.0908**	0.2727 ± 0.1052	0.2485 ± 0.3321
		$\Delta\theta$	**10.6266 ± 2.8185**	15.5265 ± 1.5682	12.9016 ± 1.7077	11.2975 ± 1.2784
PurR50	76	*Sn*	0.2633 ± 0.1372	0.1591 ± 0.0535	0.4523 ± 0.1815	**0.8570 ± 0.0634**
		Sp	0.3143 ± 0.1534	0.1857 ± 0.0595	0.5238 ± 0.2073	**1.0000 ± 0.0000**
		$\Delta\theta$	44.6861 ± 11.0815	74.1842 ± 13.4898	55.7596 ± 7.1127	**41.4494 ± 8.5976**
SoxS50	78	*Sn*	0.1495 ± 0.1135	0.1161 ± 0.0527	0.4527 ± 0.2263	**0.8807 ± 0.0455**

		Sp	0.1680 ± 0.1343	0.1344 ± 0.0585	0.5210 ± 0.2578	**1.0000 ± 0.0000**
		$\Delta\theta$	**33.2834** ± 10.7088	74.1765 ± 13.6115	58.4804 ± 17.9849	35.8106 ± **1.9104**
TyrR50	82	*Sn*	0.1750 ± 0.1061	0.1313 ± 0.0470	0.3829 ± 0.1387	**0.8607 ± 0.0612**
		Sp	0.2016 ± 0.1236	0.1512 ± 0.0501	0.4412 ± 0.1594	**1.0000 ± 0.0000**
		$\Delta\theta$	45.3958 ± 24.3895	90.8676 ± 17.3854	70.6216 ± 15.4303	**43.1779 ± 5.8466**

Πίνακας 5.4: Αποτελέσματα για την $2^{η}$ Κατηγορία Αλυσίδων DNA με Αρχικοποίηση των K-κέντρων.

Σύνολο αλυσίδων	Μήκος Μοτίβου	Μέτρο επίδοσης	Μέθοδος			
			Ext	*Bound*	*EM*	*GS*
AraC50	17	*Sn*	0.2935 ± 0.2484	0.1353 ± 0.0603	0.3824 ± 0.2855	**0.5468 ± 0.3609**
		Sp	0.3523 ± 0.2769	0.1667 ± 0.0853	0.4583 ± 0.3029	**0.6591 ± 0.4057**
		$\Delta\theta$	29.9858 ± 19.5623	51.4896 ± 38.7265	34.2904 ± 24.9755	**28.9093 ± 17.5851**
CytR50	40	*Sn*	0.1861 ± 0.1161	0.1842 ± 0.1146	0.2869 ± 0.1087	**0.3815 ± 0.2221**
		Sp	0.2917 ± 0.1881	0.2833 ± 0.1801	0.4583 ± 0.1881	**0.5333 ± 0.3798**

		$\Delta\theta$	30.4243 ± 4.9087	37.1958 ± 7.2420	29.0173 ± 5.3006	**25.2636 ± 1.3938**
DnaA50	9	*Sn*	0.0618 ± 0.0753	**0.1662 ± 0.1795**	0.1294 ± 0.1221	0.1270 ± 0.1600
		Sp	0.0936 ± 0.1338	**0.2217 ± 0.2402**	0.1872 ± 0.1951	0.2020 ± 0.2870
		$\Delta\theta$	10.1045 ± 2.0438	9.9606 ± 2.1013	**8.1862 ± 1.4574**	8.2435 ± 1.2662
LexA50	20	*Sn*	0.1515 ± 0.1328	0.2258 ± 0.1682	**0.2392 ± 0.1715**	0.1497 ± 0.1591
		Sp	0.2000 ± 0.1969	0.2765 ± 0.2047	**0.3353 ± 0.2290**	0.2529 ± 0.3165
		$\Delta\theta$	17.5234 ± 2.8432	20.8174 ± 2.9085	16.4511 ± 4.0873	**14.1510 ± 1.4880**
Nac50	15	*Sn*	0.0958 ± 0.1740	0.0969 ± 0.0782	0.1444 ± 0.1953	**0.2695 ± 0.3485**
		Sp	0.1111 ± 0.1999	0.1263 ± 0.1045	0.1768 ± 0.2292	**0.3131 ± 0.4124**
		$\Delta\theta$	17.0373 ± 18.4056	24.7345 ± 30.2274	17.8999 ± 20.9405	**16.7356 ± 14.4369**
Fur50	19	*Sn*	0.1043 ± 0.1917	0.1304 ± 0.1431	0.1563 ± 0.2195	**0.2623 ± 0.2866**
		Sp	0.1233 ± 0.2235	0.1867 ± 0.1866	0.2033 ± 0.2702	**0.3900 ± 0.4045**
		$\Delta\theta$	17.1278 ± 15.2889	22.2681 ± 20.6106	16.6065 ± 18.2688	**14.9189 ± 9.4452**
FruR50	14	*Sn*	0.1252 ± 0.1514	0.1763 ± 0.2544	**0.2507 ± 0.1898**	0.0922 ± 0.1557
		Sp	0.1364 ± 0.2014	0.2091 ± 0.3208	**0.3727 ± 0.2412**	0.1273 ± 0.2370

		$\Delta\theta$	13.9357 ± **1.2889**	16.3435 ± 3.5486	**13.7427** ± 2.3336	13.8446 ± 1.6542
GlpR50	80	*Sn*	0.5197 ± 0.1066	0.0898 ± 0.0049	0.5545 ± 0.1831	**0.8772 ± 0.0580**
		Sp	0.6015 ± 0.1181	0.1053 ± 0.0000	0.6316 ± 0.2044	**1.0000 ± 0.0000**
		$\Delta\theta$	47.3641 ± 5.7794	86.7363 ± 7.1266	55.3735 ± 5.4311	**44.7349 ± 3.6804**
NtrC50	15	*Sn*	0.0146 ± 0.0357	0.1204 ± 0.0840	0.0314 ± 0.0266	**0.2606 ± 0.2399**
		Sp	0.0202 ± 0.0670	0.1818 ± 0.1340	0.0505 ± 0.0580	**0.3434 ± 0.3041**
		$\Delta\theta$	12.2244 ± 0.0000	13.4295 ± 2.2297	**12.3127 ± 0.1953**	14.9958 ± 2.1029
OmpR50	10	*Sn*	0.0413 ± 0.0515	0.0603 ± 0.0553	0.0861 ± 0.0672	0.0982 ± 0.1605
		Sp	0.0788 ± 0.1108	0.0849 ± 0.0899	0.1394 ± 0.1052	**0.1636 ± 0.2881**
		$\Delta\theta$	6.3507 ± 3.6617	7.6474 ± 3.0343	5.0984 ± 1.7015	**5.5324 ± 0.3424**
OxyR50	45	*Sn*	0.3236 ± 0.0859	0.3054 ± 0.0866	0.4133 ± 0.0847	**0.4853 ± 0.1645**
		Sp	0.4545 ± 0.1161	0.4141 ± 0.1414	0.6061 ± 0.1038	**0.6566 ± 0.2917**
		$\Delta\theta$	32.2243 ± 1.6119	38.2263 ± 2.7484	30.9829 ± 3.0425	**29.6489 ± 1.3029**
PhoB50	17	*Sn*	0.0845 ± 0.0982	0.1445 ± 0.0761	0.1149 ± 0.1028	**0.1536 ± 0.1804**
		Sp	0.1030 ± 0.1345	0.1879 ± 0.1148	0.1636 ± 0.1441	**0.2485 ± 0.2953**

		$\Delta\theta$	12.2940 ± 4.4816	13.5685 ± 3.7323	**10.9444** ± 3.0032	11.3703 ± **1.3055**
PurR50	76	*Sn*	0.5036 ± 0.1031	0.1211 ± 0.0253	0.6024 ± 0.1806	**0.8728 ± 0.0700**
		Sp	0.5762 ± 0.1066	0.1428 ± 0.0242	0.6905 ± 0.1915	**1.0000 ± 0.0000**
		$\Delta\theta$	43.0652 ± 3.4941	85.9165 ± 12.8569	51.2687 ± 2.7665	**43.0052 ± 2.2961**
SoxS50	78	*Sn*	0.4337 ± 0.1233	0.1077 ± 0.0296	0.5478 ± 0.1544	**0.8644 ± 0.0417**
		Sp	0.4958 ± 0.1398	0.1260 ± 0.0315	0.6345 ± 0.1790	**1.0000 ± 0.0000**
		$\Delta\theta$	38.9962 ± 3.8610	77.7407 ± 10.2243	47.5586 ± 5.0287	**36.5064 ± 1.4281**
TyrR50	82	*Sn*	0.4316 ± 0.1196	0.1064 ± 0.0285	0.5297 ± 0.1606	**0.8818 ± 0.0391**
		Sp	0.5042 ± 0.1398	0.1260 ± 0.0315	0.6135 ± 0.1856	**1.0000 ± 0.0000**
		$\Delta\theta$	47.6182 ± 5.5064	89.4206 ± 10.2391	55.7193 ± 4.9048	**43.1748 ± 6.2095**

Τα συμπεράσματα που προκύπτουν με βάση τους πίνακες 5.1 και 5.3 (τυχαία αρχικοποίηση του πίνακα θ) είναι εξαιρετικά ενδιαφέροντα όσον αφορά τις επιδόσεις των μεθόδων EM, Ext, Bound και GS. Όπως είναι φανερό, ο αλγόριθμος EM, κατά την εφαρμογή του σε πραγματικά δεδομένα, παρουσιάζει τις χαμηλότερες επιδόσεις στα ποσοστά επιτυχίας Sn και Sp. Εξαιρετικά ενδιαφέρον είναι το γεγονός ότι η μέθοδος δειγματοληψίας του Gibbs δίνει τα καλύτερα μέτρα επίδοσης στο μεγαλύτερο ποσοστό των πραγματικών συνόλων δεδομένων που εξετάσαμε. Όπως είναι αναμενόμενο στις περιπτώσεις, όπου το ζητούμενο μοτίβο είναι αρκετά μεγάλο (ArcA20, CytR20, OxyR20, GlpR50, PurR50, SoxS50, TyrR50), ο τελευταίος αλγόριθμος (GS) υπερισχύει πάντα έναντι των υπολοίπων (καθώς έχει μεγαλύτερη πιθανότητα να εντοπίσει τις θέσεις μοτίβου σε κάθε συμβολοσειρά (Sn, Sp) σε σχέση με τις περιπτώσεις που το μοτίβο είναι περιορισμένου μήκους και στις οποίες υστερεί σημαντικά), ενώ ακολουθούν ο αλγόριθμος του οριοθετούμενου πίνακα (Bound) και ο EM. Το ερώτημα που τίθεται είναι για ποιο λόγο ο αλγόριθμος του εκτεταμένου πίνακα (Ext) υστερεί συγκρινόμενη με τις άλλες μεθόδους. Η συμπεριφορά αυτή θα μπορούσε να θεωρηθεί ως αναμενόμενη, στην περίπτωση ανίχνευσης ενός μοτίβου μεγάλου μήκους, αν αναλογισθεί κανείς τον εκτεταμένο αριθμό στηλών του πίνακα της μεθόδου Ext. Το μειονέκτημα αυτό, λοιπόν, την καθιστά αδύναμη στην επιλογή μεταξύ ενός μεγάλου πλήθους πιθανών μοτίβων, εκείνου που «ταιριάζει» καλύτερα σε αυτό που αναζητούμε. Λόγω του ότι η μέθοδος δειγματοληψίας του Gibbs δεν αντιμετωπίζει ένα τέτοιο πρόβλημα και εφόσον εμφανίζει τις μικρότερες τιμές στην ποσότητα $\Delta\theta$, προσεγγίζει καλύτερα από τις υπόλοιπες μεθόδους και τον πίνακα αναπαράστασης του ζητούμενου μοτίβου. Βέβαια, η μέθοδος του εκτεταμένου πίνακα (Ext) υπερέχει από όλες τις άλλες, όταν το μήκος του μοτίβου είναι περιορισμένο (μέχρι 18 σύμβολα). Επιπλέον, κάτι που παρατηρήθηκε και θα μπορούσε να χαρακτηριστεί ως παράδοξο όσον αφορά την ποσότητα $\Delta\theta$ είναι ότι ο EM δίνει χαμηλότερες τιμές κατά μέσο όρο (δηλαδή έχουμε καλύτερη προσέγγιση του πίνακα αναπαράστασης θ) από τις μεθόδους Ext και Bound. Μια λογική εξήγηση, στην οποία αποδίδουμε το γεγονός αυτό, έγκειται στο ότι τα σύνολα των συμβολοσειρών, στα οποία αναζητούμε το μοτίβο, παρουσιάζουν μεγάλο ποσοστό μετάλλαξης. Μάλιστα, το ποσοστό αυτό είναι μεγαλύτερο από 0.4 που είχαμε χρησιμοποιήσει στα (τεχνητά) πειράματα που προηγήθηκαν.

Εξετάζοντας τους πίνακες 5.2 και 5.4, αντιλαμβανόμαστε ότι τα αποτελέσματα εμφανίζουν διαφορά σε κάποια σημεία, αρχικοποιώντας τον πίνακα αναπαράστασης του μοτίβου με τον αλγόριθμο των K-κέντρων. Έτσι, στην περίπτωση που έχουμε μικρό μήκος

μοτίβου (συγκεκριμένα ≤ 16) δεν ευνοείται απαραίτητα η μέθοδος του εκτεταμένου πίνακα (Ext), αλλά δίνει καλύτερα αποτελέσματα στα μέτρα *Sn* και *Sp* από τον ΕΜ. Σε ό,τι αφορά, τώρα, το μέτρο επίδοσης $\Delta\theta$, καθώς και τη συμπεριφορά των μεθόδων για μεγάλου μήκους μοτίβα, οι παρατηρήσεις μας δεν διαφοροποιούνται. Συνεπώς, το συμπέρασμα που προκύπτει, και δεν είχε γίνει εμφανές μέσα από τα πειράματα των προηγούμενων ενοτήτων, είναι ότι η αρχικοποίηση του πίνακα θ και το μήκος του μοτίβου επηρεάζουν άμεσα το αποτέλεσμα που πρόκειται να δώσει και ο αλγόριθμος Ext εκτός από τον ΕΜ.

ΚΕΦΑΛΑΙΟ 6. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα διατριβή, μελετήσαμε το πρόβλημα ανίχνευσης μοτίβων καθορισμένου μήκους μέσα σε ένα σύνολο συμβολικών ακολουθιών και κατά κύριο λόγο, επιχειρήσαμε να το επιλύσουμε ανάγοντάς το σε πρόβλημα εκτίμησης παραμέτρων. Έτσι, παρουσιάσαμε, αρχικά, δύο γνωστές επαναληπτικές μεθόδους που χρησιμοποιούνται ευρέως και στοχεύουν στον προσδιορισμό των παραμέτρων ενός μοντέλου, όπως είναι ο ΕΜ και ο Δειγματολήπτης Gibbs (Gibbs Sampling). Στην προσπάθειά μας να προσεγγίσουμε το ίδιο πρόβλημα κάτω από διαφορετική σκοπιά, περιγράψαμε τον αλγόριθμο κατασκευής Πιθανοτικών Δέντρων Προθεμάτων (PST-Build), μία νέα και πολλά υποσχόμενη στατιστική μέθοδο ανίχνευσης μοτίβων, η οποία λειτουργεί αυξητικά αναθέτοντας ένα μέτρο βαρύτητας σε κάθε κόμβο του δέντρου που κατασκευάζει. Η αρχική, τώρα, τυχαία επιλογή των παραμέτρων (και κυρίως της βασικής παραμέτρου, δηλαδή του πίνακα αναπαράστασης του μοτίβου) που χρησιμοποιούν οι δύο πρώτες μέθοδοι, μας οδήγησε στη μελέτη δύο μεθόδων ομαδοποίησης, του Συνθετικού αλγορίθμου και του αλγορίθμου των Κ-κέντρων. Αυτό συνέβη, προκειμένου να παραχθεί, μέσω των δύο τελευταίων, μια αρχική προσέγγιση του πίνακα αναπαράστασης του μοτίβου, καθώς και για την μεταξύ τους σύγκριση κατά την διεξαγωγή των πειραμάτων που πραγματοποιήθηκαν.

Η κυρίαρχη, βέβαια, συμβολή της διατριβής στην επίλυση του εξεταζόμενου προβλήματος βασίζεται στις δύο νέες μεθόδους που προτείναμε για τον εντοπισμό του μοτίβου, τις οποίες ορίσαμε ως μέθοδο του Εκτεταμένου (Extended array) και του Οριοθετούμενου πίνακα (Bounded array). Αυτές χρησιμοποιούν την ιδέα κατασκευής του ΕΜ, υπό την έννοια ότι έχουν ως στόχο τη μεγιστοποίηση της λογαριθμικής συνάρτησης πιθανοφάνειας που ορίζεται από το σύνολο των παρατηρήσεων. Το πλεονέκτημά τους, όμως, οφείλεται στο γεγονός ότι την διευρύνουν, καθώς χρησιμοποιούν ένα πληρέστερο παραμετρικό μοντέλο από τον ΕΜ και μεγαλύτερο μέρος της πληροφορίας που παρέχεται μέσω των εκάστοτε παρατηρήσεων.

Όπως, λοιπόν, προκύπτει από το προηγούμενο κεφάλαιο, οι δύο πρωτοεμφανιζόμενες μέθοδοι παρουσιάζουν καλύτερες επιδόσεις από τους αλγορίθμους ΕΜ και Gibbs Sampling στην περίπτωση των τεχνητών δεδομένων που χρησιμοποιήσαμε και κυρίως, κατά την αναζήτηση ενός μοτίβου με επαναλαμβανόμενα σύμβολα. Το γεγονός αυτό οφείλεται ακριβώς στο ότι εκμεταλλεύονται μεγαλύτερο ποσοστό της κρυμμένης πληροφορίας που εμπεριέχει το εκάστοτε σύνολο εκπαίδευσης. Επιπλέον, αξιοσημείωτο είναι το ότι στην περίπτωση αρχικοποίησης της βασικής παραμέτρου με τον αλγόριθμο των Κ-κέντρων (ο οποίος δεν αποτελεί και την ιδανικότερη μέθοδο ομαδοποίησης δεδομένων), οι μέθοδοι του Εκτεταμένου και του Οριοθετούμενου πίνακα υπερισχύουν με σαφή διαφορά έναντι των ΕΜ και Δειγματολήπτη Gibbs στα μέτρα επίδοσης που χρησιμοποιήσαμε. Άρα, δεν εξαρτώνται σε τόσο μεγάλο βαθμό από την αρχικοποίηση της παραμέτρου που θέλουμε να προσεγγίσουμε όσο ο ΕΜ.

Από την άλλη πλευρά, κατά την πειραματική μελέτη σε πραγματικά σύνολα δεδομένων έγιναν εμφανή και ορισμένα από τα αδύνατα σημεία των ίδιων μεθόδων. Συγκεκριμένα, ο αλγόριθμος του Εκτεταμένου πίνακα δεν εντοπίζει την κρυμμένη πληροφορία στις περιπτώσεις μοτίβων πολύ μεγάλου μήκους, ενώ εξαρτάται και από την αρχικοποίηση (όχι, όμως, όσο ο ΕΜ) σε αντίθετη με αυτόν του Οριοθετούμενου πίνακα ο οποίος παρουσιάζει μια σταθερή εικόνα. Παρ' όλα αυτά, στις περιπτώσεις μοτίβων περιορισμένου μήκους φαίνεται πως είναι ο νικητής.

Όσον αφορά, τώρα, τις μεθόδους αρχικοποίησης, προκύπτει το συμπέρασμα ότι ο Συνθετικός αλγόριθμος (Agglomerative) πλεονεκτεί έναντι αυτού των Κ-κέντρων, εφόσον προσφέρει καλύτερες προϋποθέσεις εντοπισμού του κρυμμένου μοτίβου από τον ΕΜ (περίπτωση τεχνητών δεδομένων). Ακόμα πιο φανερή είναι η αδυναμία του τελευταίου αλγορίθμου κατά την μελέτη των πραγματικών δεδομένων, όπου υπάρχει μεγάλο ποσοστό μετάλλαξης του μοτίβου, αφού η τυχαία αρχικοποίηση παρέχει καλύτερα αποτελέσματα, γεγονός που αποδεικνύεται περίτρανα από την υπερίσχυση του Δειγματολήπτη Gibbs, ο οποίος υστερούσε σημαντικά έναντι των υπολοίπων στα πειράματα πάνω σε τεχνητά δεδομένα.

Με βάση τις παραπάνω παρατηρήσεις, ανακύπτουν σημαντικά θέματα που θα μπορούσαν να αποτελέσουν το έναυσμα περαιτέρω μελέτης του προβλήματος που εξετάσαμε. Ένα από αυτά εντοπίζεται στην ανεύρεση αποτελεσματικότερων κριτηρίων επιλογής, κατά την

ανίχνευση ενός μοτίβου μεγάλου μήκους, της θέσης έναρξης του μοτίβου στην περίπτωση της μεθόδου του Εκτεταμένου πίνακα.

Παρόμοια, στην περίπτωση της μεθόδου του Οριοθετούμενου πίνακα παραμένει αναπάντητο το ερώτημα της ύπαρξης ή μη κάποιου καλύτερου κριτηρίου επιλογής της πραγματικής θέσης του μοτίβου από αυτό που χρησιμοποιήσαμε.

Τέλος, ένα εξίσου σημαντικό θέμα που θα ήταν ενδιαφέρον να λάβει απάντησης είναι ο σχεδιασμός μιας μεθόδου ομαδοποίησης, ικανής να δώσει μια ικανοποιητική αρχική προσέγγιση της βασικής παραμέτρου (πίνακας θ) του προβλήματος, όταν υπάρχει στα δεδομένα μεγάλο ποσοστό μετάλλαξης (συγκεκριμένα μεγαλύτερο από 0.4).

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] T.L. Bailey and C. Elkan. Fitting a mixture model by E-M to discover motifs in biopolymers, UCSD Technical Report CS94-351.

[2] T.L. Bailey and C. Elkan. Unsupervised Learning of multiple motifs in Biopolymers using Expectation Maximization, Machine Learning, 21:51-83, 1995.

[3] G. Bejerano and G. Yona. Variations on probabilistic suffix trees: statistical modeling and prediction of protein families. Bioinformatics, 17(1):23-43, 2001.

[4] J.A. Bilmes. A gentle tutorial of the EM Algorithm and its application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, International CS Institute Berkeley CA, Technical Report, University of Berkeley, ICSI-TR-97-021, 1998.

[5] K. Blekas, D.I. Fotiadis and A. Likas. Greedy mixture learning for multiple motif discovering in biological sequences. Bioinformatics, vol. 19(5):607-617, 2003.

[6] K. Blekas and A. Likas. Incremental Mixture Learning for Clustering Discrete Data. 3rd Hellenic Conference on Artificial Intelligence, 2004.

[7] K. Blekas. A Mixture Model based Markov Random Field for Discovering Patterns in Sequences. Panhellenic Conference in Artificial Intelligence (SETN-2006) Heraclion, Greece, May 2006, Lecture Notes in Artificial Intelligence, vol. 3955, pp. 25 – 34, 2006.

[8] B. Brejova, C. DiMarco, T. Vinar, S. Romero Hidalgo, G. Holguin and C. Patten. Finding patterns in Biological Sequences. Project Report for CS7989, University of Waterloo, 2000.

[9] S.P. Brooks. Markov chain Monte Carlo method and its application. The Statistician, Vol. 47, Part 1, pp. 69-100, 1998.

[10] A. Dempster, N. Laird and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39(1), 1-22, 1977.

[11] J. Hu, B. Li and D. Kihara. (2005) Limitations and Potentials of Current Motif Discovery Algorithms. Nucleic Acids Res. Vol. 33(15), pp. 4899-4913, 2005.

[12] J. Hu, Y. D. Yang and D. Kihara. EMD: an ensemble algorithm for discovering regulatory motifs in DNA sequences. BMC Bionformatics vol. 7:342, 2006.

[13] C. Kermorvant and P. Dupont. Improved Smoothing for Probabilistic Suffix Trees Seen as Variable Order Markov Chains. Springer – Verlag Berlin Heidelberg, 2002.

[14] M. Kuroda and M. Sakakihara. Accelerating the convergence of the EM algorithm using the vector ε algorithm. Computational Statistics and Data Analysis, 51:1549-1561, 2006.

[15] C.E. Laurence, S.F. Altschul, M.S. Boguski, J.S. Liu, A.F. Neuwold and J.C. Wooton. Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment, Science 262:208-214.

[16] B. Leibe, K. Mikolajczyk and B. Schiele. Efficient clustering and matching for object class recognition, *ICCV*, pages 1792–1799, 2005.

[17] T. Li, S. Ma and M. Ogihara. Entropy-Based Criterion in Categorical Clustering, Proceedings of the 21st International Conference on Machine Learning, Banff, Canada, 2004.

[18] X. Liu, D.L. Brutlag and J.S. Liu. Bioprospector: Discovering Conserved DNA Motifs in Upstream Regulatory Regions of Co-expressed Genes, Pacific Symposium on Biocomputing, 6:127-138, 2001.

[19] M. Meila and D. Hecherman. An experimental comparison of model-basedclustering methods. Machine Learning, 42:9-29, 2001.

[20] I. Rigoutsos and A. Floratos. Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. Bionformatics, 14, 55-67, 1998.

[21] E.C. Rouchka. A Brief Overview of Gibbs Sampling. IBC Statistics Study Group, Washington University, Institute for Biomedical Computing, 1997.

[22] H.T. Wareham, T. Jiang and X. Zhang. Stochastic Heuristic Algorithms for Target Motif Identification. In Proceedings of the Fifth Pacific Symposium, pp. 389-400. World Scientific Press, Singapore, 2000.

[23] M. Welling and K. Kurihara. Bayesian K-Means as a "Maximization-Expectation" Algorithm. In Siam Conference on Data Mining, 2006.

[24] X. Wu, B. Wang, C. Song and J. Cheng. A Combined Model and a Varied Gibbs Sampling Algorithm Used for Motif Discovery. Australian Computer Society, Inc. 2004.

[25] X. Zhou, X. Zhang, X. Hu. Semantic Smoothing of Document Models for Agglomerative Clustering. 20th International Joint Conference on Artificial Intelligence (IJCAI), Jan. 6-12, 2007.

ΒΙΟΓΡΑΦΙΚΟ

Η Βουδιγάρη Έλλη εισήχθη στο Τμήμα Μαθηματικών του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών το 2000 και έλαβε το Πτυχίο Μαθηματικών το 2004. Από τον Φεβρουάριο του 2005 είναι Μεταπτυχιακή φοιτήτρια στο Τμήμα Πληροφορικής του Πανεπιστημίου Ιωαννίνων.

Στα επιστημονικά της ενδιαφέροντα συμπεριλαμβάνονται προβλήματα Μηχανικής Μάθησης, Αναγνώρισης Προτύπων και αριθμητικές μέθοδοι για την επίλυση Διαφορικών Εξισώσεων.