Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής
και Υπολογιστών

Μέτρα της οργάνωσης και της ποιότητας για το υβριδικό νευρωνικό δίκτυο Self-Organizing Hidden Markov Model Map

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Γιάννης Ν. Παπανικολάου

Επιβλέπων : Ανδρέας – Γεώργιος Ν. Σταφυλοπάτης
Καθηγητής ΕΜΠ

Αθήνα, Ιούλιος 2008

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής
και Υπολογιστών

Μέτρα της οργάνωσης και της ποιότητας για το υβριδικό νευρωνικό δίκτυο Self-Organizing Hidden Markov Model Map

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Γιάννης Ν. Παπανικολάου

Επιβλέπων : Ανδρέας – Γεώργιος Ν. Σταφυλοπάτης
Καθηγητής ΕΜΠ

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 23η Ιουλίου 2008

.....	.....	.....
Ανδρέας – Γεώργιος Ν. Σταφυλοπάτης Καθηγητής ΕΜΠ	Στέφανος Κόλλιας Καθηγητής ΕΜΠ	Παναγιώτης Τσανάκας Καθηγητής ΕΜΠ

Αθήνα, Ιούλιος 2008

.....

Γιάννης Ν. Παπανικολάου

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Copyright © Γιάννης Ν. Παπανικολάου, 2008.
Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περίληψη

Αντικείμενο της παρούσας διπλωματικής εργασίας αποτελεί η ανάπτυξη και η μελέτη μέτρων της οργάνωσης και της ποιότητας για το υβριδικό νευρωνικό δίκτυο Self-Organizing Hidden Markov Model Map (SOHMMM). Αρχικά, μελετάται μια πλειάδα ήδη υπάρχοντων μέτρων για αλγόριθμους εύρεσης συστάδων. Στη συνέχεια, με βάση αυτά, αναπτύσσονται νέα μέτρα οργάνωσης εξειδικευμένα για το SOHMMM και τελικά διερευνάται η επίδοση και η λειτουργικότητά τους μέσω προσομοιώσεων με τεχνητά και πραγματικά σύνολα δεδομένων. Στόχος της ανάπτυξης των μέτρων αυτών είναι η παρακολούθηση της διαδικασίας εκπαίδευσης του δικτύου (για παράδειγμα σε λογικές πρόωρου τερματισμού) καθώς και ο προσδιορισμός μιας βέλτιστης παραμετροποίησης για το δίκτυο με βάση το συγκεκριμένο σύνολο δεδομένων.

Στο Κεφάλαιο 1 πραγματοποιείται μια εισαγωγή στη μελέτη που έγινε καθώς και σε κάποιες βασικές έννοιες. Στα Κεφάλαια 2 και 3 παρουσιάζονται αντίστοιχα ο αυτο-οργανούμενος χάρτης (SOM) και το Hidden Markov Model που αποτελούν τα συστατικά μέρη του υβριδικού νευρωνικού δικτύου SOHMMM. Το τελευταίο παρουσιάζεται στο Κεφάλαιο 4. Στο Κεφάλαιο 5 γίνεται μια αναλυτική περιγραφή των μέτρων οργάνωσης που μελετήθηκαν. Παρουσιάζονται τα μέτρα με παράλληλη αναφορά στην αρχική μορφή και χρήση τους ενώ στη συνέχεια αναφέρονται οι τροποποιήσεις και προσαρμογές που έγιναν προκειμένου να χρησιμοποιηθούν στο SOHMMM. Για κάθε ένα από τα μέτρα δίνεται και η αναμενόμενη συμπεριφορά του κατά τη διάρκεια της διαδικασίας εκπαίδευσης ή ανάλογα με διαφορετικές παραμετροποιήσεις. Το Κεφάλαιο 6 αφορά στις πραγματοποιηθείσες προσομοιώσεις. Παρουσιάζονται τα αποτελέσματα και εξάγονται κάποια συμπεράσματα σχετικά με τις επιδόσεις και τη λειτουργικότητα του κάθε μέτρου.

Λέξεις Κλειδιά

αυτο-οργανούμενα νευρωνικά δίκτυα, Self-Organizing Hidden Markov Model Map, εύρεση συστάδων ακολουθιακών δεδομένων, μέτρα οργάνωσης, πρόωρος τερματισμός, βέλτιστη παραμετροποίηση, εντροπία ενεργοποίησης νευρώνα, σφάλμα κβαντισμού, τοπογραφικό σφάλμα, τοπογραφικό γινόμενο, παραμόρφωση, C-measure, Calinski-Harabasz, Demartines-Herault.

Abstract

The purpose of the present diploma thesis is the development and study of organization and quality measures for the Self-Organizing Hidden Markov Model Map (SOHMMM). Initially, we study a number of such already existing measures for clustering algorithms. Next, based on these, we develop new organization measures, specialized for the SOHMMM and finally we investigate their performance and functionality via simulations with artificial and real datasets. The final goal of the development of the aforementioned measures is to be able to follow the evolution of the training process (e.g. in an early stopping approach) as well as to determine the optimal parametrization for a network, given the dataset.

In Chapter 1 we are making an introduction to the study that was done as well as to some basic notions of our work. Chapter 2 and 3 deal with the Self-Organizing Map (SOM) and the Hidden Markov Model (HMM) respectively. The SOHMMM, which is presented in Chapter 4, is essentially based on these two models. In Chapter 5, we describe thoroughly the organization measures that were studied. We present the measures, referring to their initial form and use and next we mention the changes and adaptations that were realized so as to be able to use them for the SOHMMM. For each one of these measures we also describe the expected behavior during the training process or for different parametrizations. Chapter 6 deals with the realized simulations. We present the results and the conclusions we reached concerning the performance and functionality of each measure

Keywords

self-organization, Self-Organizing Hidden Markov Model Map, sequence clustering, organization measures, quality measures, early stopping, optimal parametrization, neuron activation entropy, quantization error, topographic error, topographic product, distortion, C-measure, Calinski-Harabasz, Demartines-Herault

Ευχαριστίες

Για την εκπόνηση της παρούσας διπλωματικής εργασίας θα ήθελα να ευχαριστήσω κατά κύριο λόγο τον καθηγητή μου κ. Α. Γ. Σταφυλοπάτη, που με την αμέριστη συμπαράσταση και τις πάντοτε καίριες και χρήσιμες συμβουλές του υπήρξε ο καθοδηγητής μου στη μελέτη αυτή. Επίσης, η εργασία αυτή δεν θα είχε έλθει εις πέρας χωρίς την πολύτιμη συμβολή και τις παρεμβάσεις του υποψήφιου διδάκτορα Χρήστου Φερλέ που βοήθησε σημαντικά στο να ξεπεραστούν τα σημαντικά προβλήματα που προέκυψαν.

Πίνακας Περιεχομένων

Κεφάλαιο 1

1 Εισαγωγή……...15

Κεφάλαιο 2

2 Αυτο-οργανούμενοι χάρτες (Self-Organizing Maps-SOM)……..17
- 2.1 Εισαγωγή……19
- 2.2 Μια ποιοτική προσέγγιση……...……19
- 2.3 Ο βασικός αλγόριθμος του αυτο-οργανούμενου χάρτη……21
- 2.4 SOM με μετρική εσωτερικού γινομένου……... 26
- 2.5 Αρχικοποίηση του αλγόριθμου του SOM……. 26
- 2.6 Βέλτιστος ρυθμός εκμάθησης…… 27
- 2.7 Επίδραση της μορφής της συνάρτησης γειτονιάς…….. 29
- 2.8 Ένα κριτήριο παραμόρφωσης για τον αλγόριθμο του SOM……. 30
- 2.9 Υποδείξεις για την κατασκευή καλών χαρτών (SOM)…….. 31
- 2.10 Οπτικοποίηση για την επίδειξη των συστάδων στο SOM……33
- 2.11 Ερμηνεία της απεικόνισης που παράγει το SOM……34

Κεφάλαιο 3

3 Hidden Markov Models-HMM……37
- 3.1 Εισαγωγή…….....39
- 3.2 Διακριτές διαδικασίες Markov……...40
- 3.3 Hidden Markov Models……..42
  - 3.3.1 Δύο απλά παραδείγματα……..42
  - 3.3.2 Παράμετροι ενός HMM……...44
  - 3.3.3 Τα τρία θεμελιώδη προβλήματα για HMMs……....46
- 3.4 Λύσεις των τριών θεμελιωδών προβλημάτων των HMMs……....47
  - 3.4.1 Πρόβλημα 1…….....47
  - 3.4.2 Πρόβλημα 2……49
  - 3.4.3. Πρόβλημα 3……51
- 3.5 Τύποι HMMs……..54
- 3.6 Σύγκριση HMMs - Ένα πιθανοτικό μέτρο απόστασης για HMMs……60

Κεφάλαιο 4

4 Το υβριδικό νευρωνικό δίκτυο SOHMMM……..63
- 4.1 Εισαγωγή…….....65
- 4.2 Bayesian Inference, εκτίμηση παραμέτρων, και επιλογή μοντέλου……..65
- 4.3 Self-Organizing Hidden Markov Model Map (SOHMMM)……..69
  - 4.3.1 Το βασικό SOHMMM……71
  - 4.3.2 Θέματα υλοποίησης για το SOHMMM……...73
- 4.4 Αλγόριθμος εκμάθησης για το SOHMMM……79

Κεφάλαιο 5

5 Μέτρα και κριτήρια της ποιότητας για το υβριδικό νευρωνικό δίκτυο SOHMMM…….83
- 5.1 Εισαγωγή…….....85
- 5.2 Γενικές έννοιες και ορισμοί……85
  - 5.2.1 Έννοιες για τα μέτρα οργάνωσης……86
  - 5.2.2 Μετρικές στον χώρο εξόδου και Τοπολογία……89

5.2.3 Απαιτήσεις για ένα μέτρο οργάνωσης…….....89
5.3 Μέτρα Οργάνωσης……91
5.3.1 Εντροπία……...91
5.3.2 Σφάλμα κβαντισμού (quantization error)……92
5.3.3 Τοπογραφικό σφάλμα (topographic error)……..93
5.3.4 C-measure……94
5.3.5 Μέτρο των Demartines – Herault……95
5.3.6 Μέτρο παραμόρφωσης (distortion measure)……...96
5.3.7 Τοπογραφικό γινόμενο (topographic product……..97
5.3.8 Το κριτήριο Kalinski – Harabasz……...100
5.3.9 Δείγματα εισόδου που αλλάζουν κόμβο αναφοράς μεταξύ δύο επαναλήψεων……..101
5.4 Μέτρα οργάνωσης που μελετήθηκαν αλλά δεν χρησιμοποιήθηκαν……101

Κεφάλαιο 6

6 Προσομοιώσεις για τα μέτρα οργάνωσης του SOHMMM……103
6.1 Εισαγωγή…….....105
6.2 Σύνολα δεδομένων……105
6.2.1 Τεχνητά σύνολα δεδομένων……...105
6.2.2 Πραγματικά σύνολα δεδομένων……107
6.3 Προσομοιώσεις……108
6.3.1 Μονοτονία εκπαίδευσης……108
6.3.1.1 Τεχνητά σύνολα δεδομένων……108
6.3.1.2 Πραγματικά σύνολα δεδομένων……..122
6.3.1.3 Συμπεράσματα……132
6.3.2 Μελέτη ως προς των αριθμό των εποχών……..135
6.3.2.1 Τεχνητά σύνολα δεδομένων……136
6.3.2.2 Globins……137
6.3.2.3 Συμπεράσματα……138
6.3.3 Μελέτη ως προς των αριθμό των κρυμμένων καταστάσεων των HMM……..139
6.3.3.1 Τεχνητά σύνολα δεδομένων……140
6.3.3.2 Πραγματικά σύνολα δεδομένων……..141
6.3.3.3 Συμπεράσματα……142
6.3.4 Μελέτη ως προς το μέγεθος του πλέγματος του SOHMMM……143
6.3.4.1 Τεχνητά σύνολα δεδομένων……144
6.3.4.2 Πραγματικά σύνολα δεδομένων……..145
6.3.4.3 Συμπεράσματα……146
6.4 Σύνοψη……..147

Βιβλιογραφία

Βιβλιογραφία…….....149

Παράρτημα

Παράρτημα Α…….....161
Παράρτημα Β…….....173

Κεφάλαιο 1
Εισαγωγή

Εισαγωγή

Στις σημερινές κοινωνίες, τα δεδομένα έχουν μια εξέχουσα σημασία σε όλες τις παραμέτρους της καθημερινότητας, των εφαρμογών και των διάφορων τομέων των κοινωνικών και οικονομικών δομών. Πρωταρχική ανάγκη αποτελεί, η οργάνωση τους και το σημαντικότερο, η ανάλυση τους προς εξαγωγή πληροφοριών. Για την επίτευξη του στόχου αυτού μια μεγάλη σειρά αλγορίθμων, εργασιών και ερευνών έχει μελετηθεί στην κατεύθυνση της οργάνωσης των δεδομένων σε συστάδες, της αφαίρεσης και τη εξαγωγής πληροφορίας.

Σε αυτό το πλαίσιο, τα ακολουθιακά δεδομένα καταλαμβάνουν ολοένα και αυξανόμενο ρόλο, από τις εφαρμογές της βιολογίας-γενετικής και της ιατρικής (ακολουθίες DNA, πρωτεϊνικές ακολουθίες, ιατρικές διαγνώσεις) ως τις τεχνολογίες αναγνώρισης φωνής, web data mining, ανάλυσης αισθητήρων ρομπότ,κλπ. Ένα πάρα πολύ χρήσιμο εργαλείο στην προσπάθεια χειρισμού των διαφόρων μορφών ακολουθιακών δεδομένων (προς το παρόν μόνο διακριτής κατανομής πιθανότητας) αποτελεί το SOHMMM.

Το SOHMMM συνιστά μία κλάση υβριδικών συστημάτων η οποία συνδυάζει τόσο σε επίπεδο αρχιτεκτονικής όσο και σε επίπεδο αλγορίθμων εκπαίδευσης/λειτουργίας τους Self-Organizing Feature Maps καθώς και τα Hidden Markov Models. Ο σχεδιασμός του αποσκοπεί στην δημιουργία και υλοποίηση ενός συστήματος υπολογιστικής νοημοσύνης με τα χαρακτηριστικά της μη-επιβλεπόμενης μάθησης και της επεξεργασίας ακολουθιακών δεδομένων (sequential data - observation sequences) υψηλών διαστάσεων και μεταβλητού μεγέθους. Το γεγονός αυτό, πρακτικά συνεπάγεται ένα σύστημα ικανό να επεξεργαστεί δεδομένα του χώρου της βιοπληροφορικής (linear DNA or protein sequences) και όχι μόνο.

Παράλληλα, στην προσπάθεια οργάνωσης των δεδομένων σε συστάδες, την αφαίρεση και την εξαγωγή πληροφοριών, εξέχοντα ρόλο παίζουν μια σειρά από μέτρα οργάνωσης και κριτήρια που είναι σε θέση να αξιολογήσουν αυτή τη διαδικασία Τα κριτήρια και μέτρα αυτά έχουν χρησιμοποιηθεί προκειμένου να είναι δυνατή μια όσο το δυνατόν πιο αντικειμενική επιλογή παραμέτρων για τους διάφορους αλγόριθμους οργάνωσης των δεδομένων καθώς και για την πρώιμο τερματισμό τους καθώς και για την παρακολούθηση της σύγκλισής τους. Στην βιβλιογραφία, έχει αναπτυχθεί ένας πολύ μεγάλος αριθμός τέτοιων μέτρων. Συνήθως, τα μέτρα αυτά ειδικεύονται σε συγκεκριμένες επιμέρους λειτουργίες και αλγόριθμους.

Στην παρούσα διπλωματική εργασία, μελετήθηκαν και αναλύθηκαν μια σειρά από τέτοια μέτρα που έχουν εφαρμοστεί σε αλγόριθμους ανίχνευσης συστάδων (κυρίως για το SOM). Στη συνέχεια, με βάση την φιλοσοφία καθενός από αυτά, αναπτύχθηκαν νέα κριτήρια, μέσω κατάλληλων μετατροπών και προσαρμογών, προκειμένου να μπορούν να ποσοτικοποιήσουν τη λειτουργία του SOHMMM. Η παρακολούθηση της διαδικασίας αυτο-οργάνωσης καθώς και η ποσοτική περιγραφή της ποιότητας τη απεικόνισης που πραγματοποιείται αποδεικνύεται ακόμα πιο σημαντική στην περίπτωση του SOHMMM, όπου αφενός δεν υπάρχει μαθηματική απόδειξη σύγκλισης (κάτι που ισχύει λόγω της καταγωγής του από το προγενέστερο SOM) ενώ παράλληλα δεν υφίσταται ένας αποτελεσματικός και μαθηματικά βασισμένος τρόπος για την επιλογή των παραμέτρων του υβριδικού δικτύου που αφορούν τόσο τις παραμέτρους για τον χάρτη (όπως μέγεθος χάρτη, αρχικό εύρος γειτονιάς, κλπ) όσο και τις παραμέτρους των HMMs (αριθμός states, αρχικό διάνυσμα π,κλπ).

Αν και μελετήθηκε ένας μεγάλος αριθμός κριτηρίων, εν τέλει επιλέχθηκαν προς τροποποίηση και προσαρμογή εννιά αντιπροσωπευτικά τέτοια μέτρα:

1. **Εντροπία.** Η εντροπία ενεργοποίησης νευρώνα, είναι ένα πολύ απλό μέγεθος από τη

θεωρία των SOM και αναμένεται να αυξάνεται για καλύτερες απεικονίσεις.

2. **Σφάλμα κβαντισμού (quantization error).** Το σφάλμα κβαντισμού, προέρχεται από την θεωρία του κβαντισμού διανυσμάτων και περιγράφει το πόσο καλά περιγράφονται τα δεδομένα από τον κόμβο αναφοράς στον οποίο έχουν αντιστοιχηθεί.
3. **Τοπογραφικό σφάλμα (topographic error).** Το τοπογραφικό σφάλμα, προερχόμενο από το SOM, περιγράφει την ομαλότητα της απεικόνισης.
4. **C-measure.** Το μέτρο αυτό, αναπτύχθηκε από τους Goodhill και Sernowski, και αποκτά αυξανόμενες τιμές για καλύτερες απεικονίσεις.
5. **Μέτρο των Demartines και Herault.** Το μέτρο αυτό αποτελεί ουσιαστικά μια συνάρτηση ενέργειας που αναμένουμε να μειώνεται για καλύτερες απεικονίσεις.
6. **Μέτρο παραμόρφωσης (distortion measure).** Ομοίως, το μέτρο παραμόρφωσης αποτελεί μια συνάρτηση ενέργειας που προσεγγίζει καλύτερα τη διαδικασία οργάνωσης του SOM.
7. **Τοπογραφικό γινόμενο (topographic product).** Το τοπογραφικό γινόμενο, αρκετά απαιτητικό υπολογιστικά, περιγράφει την ποιότητα της απεικόνισης, λαμβάνοντας την τιμή 0 για τέλεια απεικόνιση. Αποτελεί ένα από τα πιο δημοφιλή μέτρα οργάνωσης για το SOM.
8. **Κριτήριο Kalinski – Harabasz.** Το κριτήριο αυτό προέρχεται από τη θεωρία των αλγορίθμων οργάνωσης δεδομένων σε συστάδες και η συνήθης χρήση του είναι στην εύρεση του πλήθους των συστάδων στις οποίες οργανώνονται τα δεδομένα. Στην εργασία μας χρησιμοποιήθηκε, με αρκετή επιτυχία, στην περιγραφή της μονοτονίας της διαδικασίας εκπαίδευσης.
9. **Δείγματα εισόδου που αλλάζουν κόμβο αναφοράς μεταξύ δύο επαναλήψεων.** Το απλό αυτό μέτρο που αναπτύξαμε, αποτελεί ουσιαστικά έναν δείκτη της διαδικασίας εκπαίδευσης:όσο η εκπαίδευση προχωρά, αναμένεται να τείνει προς το 0.

Στη συνέχεια, μέσω τριών τεχνητών και δύο πραγματικών συνόλων δεδομένων (πρωτεϊνικές ακολουθίες) διερευνήθηκε και αναλύθηκε η λειτουργικότητα και η απόδοση των κριτηρίων αυτών στην παρακολούθηση της διαδικασίας εκπαίδευσης του συστήματος καθώς και στην εύρεση των βέλτιστων παραμέτρων για την εκπαίδευση του δικτύου.

Κεφάλαιο 2
Αυτο-οργανούμενοι χάρτες (Self-Organizing Maps-SOM)

2.1 Εισαγωγή

Ο αυτο-οργανούμενος χάρτης του Kohonen ή SOM, όπως θα αναφέρεται από εδώ και πέρα χάριν απλότητας, αποτελεί ένα νέο, αποτελεσματικό εργαλείο λογισμικού για την **οπτικοποίηση** δεδομένων μεγάλων διαστάσεων. Στη βασική του μορφή, κατασκευάζει ένα γράφο ομοιότητας των δεδομένων εισόδου. Μετατρέπει τις μη γραμμικές στατιστικές σχέσεις μεταξύ των δεδομένων μεγάλης διάστασης σε απλές γεωμετρικές σχέσεις των σημείων απεικόνισής τους. Τα σημεία αυτά οργανώνονται σε έναν χαμηλής διάστασης χάρτη, συνήθως ένα δισδιάστατο πλέγμα κόμβων. Επειδή ακριβώς το SOM συμπυκνώνει πληροφορία ενώ διατηρεί τις πιο σημαντικές τοπολογικές και μετρικές συσχετίσεις των δεδομένων μπορεί να υποστηριχθεί ότι παράγει ένα είδος **αφαίρεσης.** *Αυτά τα δύο χαρακτηριστικά, η οπτικοποίηση και η αφαίρεση σε συνδυασμό με την ομαδοποίηση προτύπων είναι δυνατό να χρησιμοποιηθούν σε μία σειρά πολύπλοκων και σύνθετων εργασιών.*

Στην παρούσα του μορφή το SOM υλοποιήθηκε από τον Teuvo Kohonen [1] το 1982. Σήμερα, περίπου 7718 εργασίες έχουν δημοσιευθεί πάνω σε αυτό [2]. Επίσης, έχουν υλοποιηθεί πακέτα λογισμικού [3] που περιέχουν τις κύριες διαδικασίες καθώς και διαγνωστικά προγράμματα και οπτικοποιήσεις για το SOM.

Το SOM θα μπορούσε να περιγραφεί τυπικά ως μια μη γραμμική, ταξινομημένη, ομαλή απεικόνιση των μεγάλης διάστασης δεδομένων εισόδου στα στοιχεία ενός συμμετρικού πίνακα χαμηλής διάστασης. Αυτή η απεικόνιση υλοποιείται με τον ακόλουθο τρόπο, που προσομοιάζει στον κλασσικό αλγόριθμο κβαντισμού του διανύσματος μάθησης (learning vector quantization - LVQ). Υποθέτουμε για λόγους απλότητας ότι το σύνολο των δεδομένων εισόδου γίνεται να οριστεί ως ένα διάνυσμα πραγματικών αριθμών $x=[\xi_1, \xi_2,...,\xi_v]$. Σε κάθε στοιχείο του SOM αντιστοιχίζουμε ένα παραμετρικό διάνυσμα $m_i = [\mu_1,\mu_2,...\mu_v]$ πραγματικών που το ονομάζουμε μοντέλο. Υποθέτοντας ένα γενικό μέτρο απόστασης (ανομοιότητας) μεταξύ των x και m_i που θα συμβολίζουμε με $d(x,m_i)$, η απεικόνιση ενός διανύσματος εισόδου x στον πίνακα του SOM ορίζεται ως το στοιχείο του πίνακα που ταιριάζει καλύτερα με το x, ή διαφορετικά που έχει τον δείκτη

$$c = \arg\min_i \{d(x, m_i)\}$$

Διαφοροποιούμενοι από τον παραδοσιακό κβαντισμό διανύσματος (VQ) σκοπός μας είναι να ορίσουμε το m_i με τέτοιο τρόπο ώστε η απεικόνιση να είναι ταξινομημένη και περιγραφική της κατανομής του x. Σε αυτό το σημείο θα πρέπει να τονιστεί ότι τα m_i δεν είναι απαραίτητο να είναι διανυσματικές μεταβλητές., αρκεί να έχει οριστεί η απόσταση $d(x,m_i)$ για όλα τα x και ένα ικανά ευρύ σύνολο από m_i.

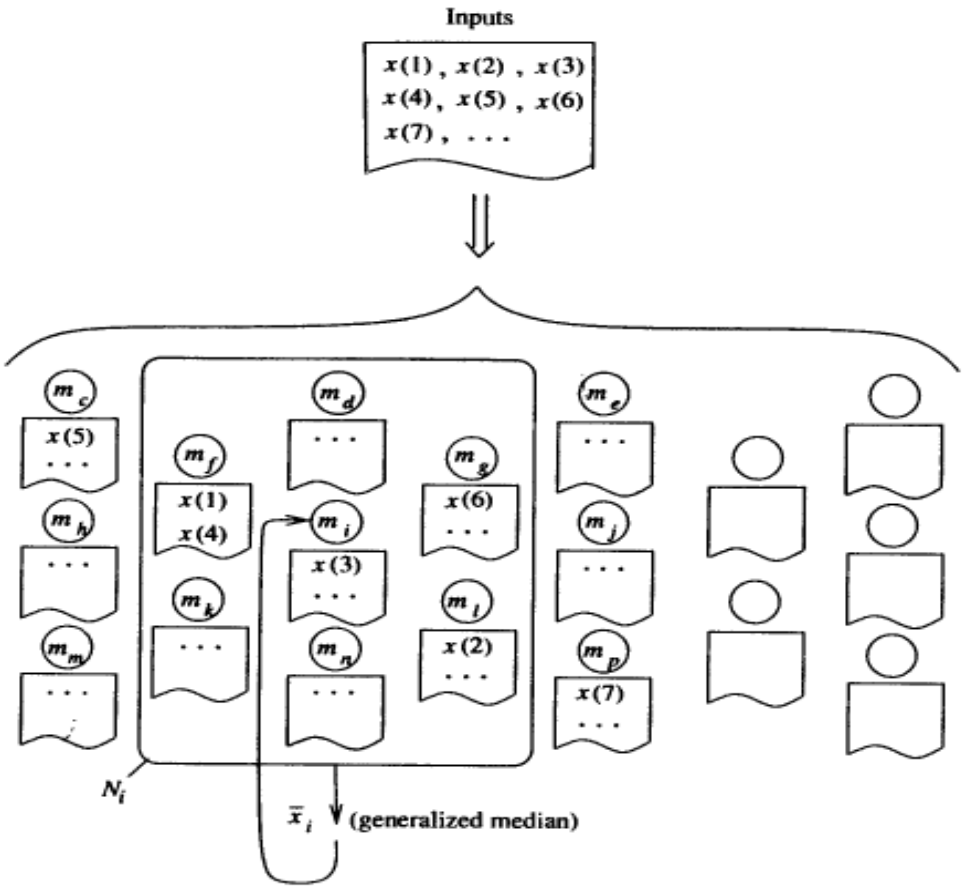
2.2 Μια ποιοτική προσέγγιση

Στην εικόνα 2.1 φαίνεται ένας δισδιάστατος πίνακας κόμβων, με κάθε κόμβο να αντιστοιχίζεται σε ένα μοντέλο m_i. Οι αρχικές τιμές για το m_i μπορούν να επιλεγούν ως τυχαίες, από τον χώρο των δεδομένων εισόδου κατά προτίμηση. Η λίστα των δεδομένων συμβολίζεται με x(t) όπου t ακέραιος δείκτης. Στο σχήμα αυτό τα m_i, x(t) μπορούν να είναι διανύσματα, συμβολοσειρές, ή σύμβολα ή κάτι ακόμα πιο γενικό. Αρχικά, συγκρίνεται κάθε x(t) με κάθε m_i αντιγράφοντας κάθε x(t) σε μια υπολίστα στο διάνυσμα-μοντέλο που είναι πιο όμοιο με αυτό, σύμφωνα με την απόσταση d. Όταν όλα τα x(t) έχουν αντιστοιχηθεί κατά αυτό τον τρόπο στους αντίστοιχους νευρώνες-κόμβους, θεωρούμε τη γειτονιά N_i γύρω από

κάθε m_i. Η έννοια της γειτονιάς περιγράφει το σύνολο των κόμβων σε μια επιλεγμένη ακτίνα γύρω από τον κόμβο i. Στη συνέχεια, στην ένωση όλων των υπολιστών στο N_i, προσδιορίζεται το πιο 'ενδιάμεσο' x_i, δηλαδή αυτό που απέχει λιγότερο από τα υπόλοιπα $x(t) \in N_i$. Το x_i αυτό καλείται στο εξής γενικός μέσος (generalized median). Αφού βρεθεί ο γενικός αυτός μέσος για κάθε κόμβο, αντικαθίσταται η παλαιά τιμή του m_i με αυτή του γενικού μέσου.

Έχει αποδειχθεί [4] πως αν η παραπάνω διαδικασία επαναληφθεί με τα x(t) και m_i ευκλείδεια διανύσματα και με μια ελαφρά διαφοροποιημένη συνάρτηση απόστασης ο αλγόριθμος αυτός συγκλίνει.

Ο αλγόριθμος του SOM που περιγράφεται παρακάτω είναι στενά συνδεδεμένος με την παραπάνω περιγραφή.

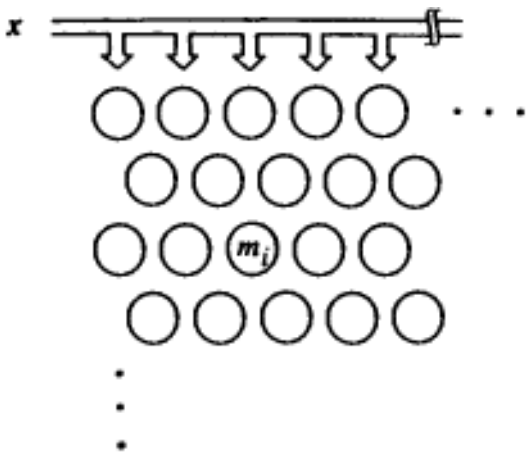


Εικόνα 2.1. Παρουσίαση της διαδικασίας όπου τα δεδομένα (δείγματα) εισόδου κατανέμονται σε υπολίστες κάτω από τα καλύτερα ταυτιζόμενα μοντέλα και επανυπολογίζονται οι τιμές των τελευταίων ως γενικευμένοι μέσοι της κάθε γειτονιάς N_i. [1]

2.3 Ο βασικός αλγόριθμος του αυτο-οργανούμενου χάρτη

Σε αυτό το σημείο προχωράμε στην περιγραφή του αλγορίθμου του SOM. Ο αλγόριθμος αυτός μπορεί να θεωρηθεί σαν μια ειδική αναδρομική οπισθοδρομούσα διαδικασία, στην οποία επεξεργαζόμαστε μόνο ένα υποσύνολο από τους κόμβους σε κάθε βήμα.

Θεωρούμε την εικόνα 2.2. Το SOM εδώ ορίζει μια απεικόνιση από τον χώρο των δεδομένων εισόδου $\mathbb{R}^n$ σε ένα δισδιάστατο πίνακα κόμβων. Σε κάθε τέτοιο κόμβο i



Εικόνα 2.2. Το πλέγμα των κόμβων-νευρώνων σε ένα δισδιάστατο SOM. [1]

αντιστοιχίζουμε ένα παραμετρικό διάνυσμα $m_i=[\mu_{i1}, \mu_{i2} .... \mu_{in}]^T \in \mathbb{R}^n$. Αρχικά το διάνυσμα m_i πρέπει να αρχικοποιηθεί, συνήθως σε μία τυχαία τιμή από το σύνολο εισόδου. Η αποτελεσματικότητα της αυτο-οργάνωσης (self-organization) έγκειται στο γεγονός ότι ακόμη και με εντελώς τυχαία αρχικά βάρη στους κόμβους, τελικά ο χάρτης αποκτά μια οργανωμένη μορφή, αναπαριστώντας τα δεδομένα.

Η μορφή του πίνακα μπορεί να οριστεί ορθογωνική, εξαγωνική ή ακόμα και μη κανονική με την εξαγωγική μορφή πάντως να είναι αποτελεσματικότερη όσον αφορά την οπτικοποίηση των δεδομένων. Στην απλούστερη περίπτωση κάθε διάνυσμα εισόδου x συνδέεται παράλληλα με όλους τους νευρώνες μέσω των βαρών m_i του κάθε ενός από αυτούς. Με μια πιο αφηρημένη προσέγγιση, θα μπορούσαμε να πούμε ότι η είσοδος x συγκρίνεται με όλα τα m_i και η τοποθεσία του καλύτερου ταιριάσματος (σύμφωνα με κάποια μετρική) ορίζεται ως η τοποθεσία της απόκρισης του δικτύου.

Έστω $x \in \mathbb{R}^N$ ένα στοχαστικό διάνυσμα δεδομένων. Θα μπορούσαμε να θεωρήσουμε πως το SOM αποτελεί μια "μη γραμμική προβολή" της συνάρτησης πυκνότητας πιθανότητας p(x) του μεγάλης διάστασης διανύσματος εισόδου x σε μια δισδιάστατη απεικόνιση. Το διάνυσμα x μπορεί να συγκριθεί με τα διανύσματα αναφοράς με χρήση διάφορων μετρικών, συνηθέστερη επιλογή όμως αποτελεί η ευκλείδεια απόσταση προκειμένου να περιγράψει την πλέον ταιριαστή μονάδα (best matching unit – BMU).

$$c = argmin\{ [x - m_i] \} \text{ ή } [x - m_c] = \min_i \{[x - m_i]\} \quad \textbf{(2.1)}$$

Κατά τη διάρκεια της εκμάθησης, δηλαδή τη διαδικασία κατά την οποία η “μη γραμμική προβολή” σχηματίζεται, οι κόμβοι που είναι 'κοντά' στο πλέγμα με βάση μια συγκεκριμένη γεωμετρική απόσταση θα ενεργοποιηθούν μαθαίνοντας από την είσοδο x. Αυτή η διαδικασία καταλήγει σε ένα φαινόμενο ομαλοποίησης (smoothing effect) των βαρών των νευρώνων της γειτονιάς αυτής, που μέσω μιας συνολικότερης διαδικασίας συνεχούς εκμάθησης οδηγεί τελικά σε μια συνολική ταξινόμηση του χάρτη.

Για την ανανέωση των βαρών χρησιμοποιείται η παρακάτω εξίσωση:

$$m_i(t+1) = m_i(t) + h_{ci}(t)\left[x(t)-m_i(t)\right] \quad \textbf{(2.2)}$$

όπου t = 0,1,2,.. είναι μια ακέραια συντεταγμένη διακριτού χρόνου.

Η συνάρτηση h(t) παίζει κεντρικό ρόλο στην διαδικασία ομαλοποίησης δρώντας ως συνάρτηση γειτονιάς (neighborhood function). Προκειμένου να συγκλίνει ο αλγόριθμος, είναι απαραίτητο να ισχύει $h_{ci} \to 0$ όταν $t \to \infty$. Συνήθως, για τη συνάρτηση γειτονιάς ισχύει

$$\boldsymbol{h_{ci}(t) = h(\|r_c(t)-r_i\|, t)} \quad \textbf{(2.3)}$$

με rc και ri να συμβολίζουν τις θέσεις των νευρώνων c και i στο δισδιάστατο πλέγμα αντίστοιχα και αύξηση της διαφοράς $\|r_c(t)-r_i\|$ να συνεπάγεται $h_{ci} \to 0$. Η μορφή της συνάρτησης γειτονιάς καθορίζει την ελαστικότητα του πλέγματος που πρόκειται να προσαρμοστεί στα δεδομένα.

Στη βιβλιογραφία, δύο είναι οι συνηθέστερες επιλογές για τη συνάρτηση γειτονιάς. Η απλούστερη θεωρεί ως γειτονιά ένα σύνολο από κόμβους γύρω από τον κόμβο-νικητή. Έστω N_c το σύνολο με τους κόμβους της γειτονιάς. Τότε θα ισχύει $h_{ci}(t)=a(t), i \in N_c$ και $h_{ci}=0$ διαφορετικά. Το a(t) καλείται ρυθμός μάθησης (learning rate) και αποκτά τιμές ανάμεσα από 0 και 1. Τόσο το εύρος της γειτονιάς N_c όσο και ο ρυθμός μάθησης a(t) συνήθως, φθίνουν γνησίως με το χρόνο.

Μια άλλη ευρέως χρησιμοποιούμενη συνάρτηση γειτονιάς είναι η συνάρτηση Gauss:

$$\boldsymbol{h_{ci}(t) = a(t) * e^{\frac{-[r_c(t)-r_i]^2}{2\sigma^2(t)}}}, \quad \textbf{(2.4)}$$

όπου το a(t) παριστάνει ξανά ρυθμό μάθησης και η παράμετρος σ, καθορίζει το εύρος του γκαουσιανού κέντρου. Οι δύο παραπάνω παράμετροι αποτελούν γνησίως φθίνουσες συναρτήσεις του χρόνου.

Στην περίπτωση όπου ο αυτοοργανούμενος χάρτης μας δεν είναι μεγάλης σχετικά διάστασης (τάξης λίγων εκατοντάδων κόμβων) η επιλογή των παραμέτρων εκπαίδευσης του δικτύου δεν είναι τόσο κρίσιμη. Για παράδειγμα, μπορεί να χρησιμοποιηθεί η απλή συνάρτηση γειτονιάς (neighborhood-set function) που περιγράψαμε παραπάνω. Ειδική μέριμνα πάντως και σε αυτή την περίπτωση πρέπει να ληφθεί σε ό, τι αφορά το εύρος της γειτονιάς. Αν το αρχικό της εύρος είναι πολύ μικρό, ο χάρτης δεν θα μπορέσει να τακτοποιηθεί συνολικά αποκτώντας μια μορφή "μωσαϊκού" με ασυνέχειες. Το φαινόμενο αυτό μπορεί να αποφευχθεί ξεκινώντας με μια μεγάλου εύρους γειτονιά η οποία συρρικνώνεται με τον χρόνο.

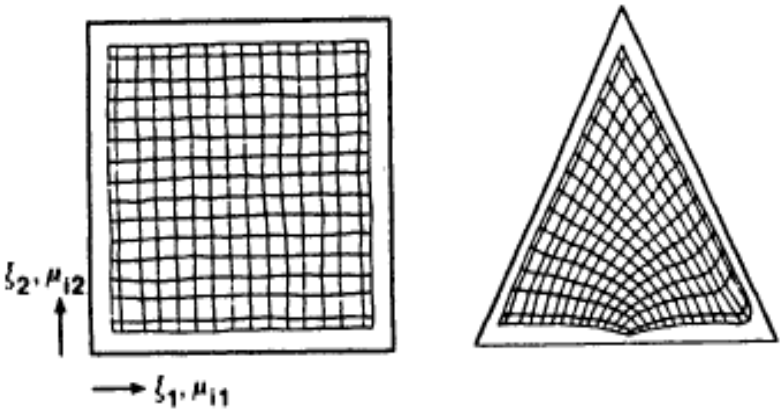
Σε ένα αρχικό στάδιο εκπαίδευσης, το λεγόμενο **ordering phase**, τόσο η συνάρτηση γειτονιάς όσο και ο ρυθμός μάθησης έχουν μεγάλες σχετικά τιμές οδηγώντας έτσι σε μεγάλες αλλαγές στην "τακτοποίηση" του χάρτη. Αυτό συμβαίνει προκειμένου να αποφευχθούν φαινόμενα "δίπλωσης" η γενικότερα μη ομαλής προσαρμογής του SOM στα δεδομένα. Σε αυτό το στάδιο το εύρος της γειτονιάς μπορεί να μειώνεται γραμμικά μέχρι την τιμή 1, δηλαδή με το σύνολο N_c να περιλαμβάνει μόνο τους άμεσους γείτονες του c στο πλέγμα. Ο ρυθμός μάθησης, εκκινώντας από μια τιμή κοντά στο 1 μπορεί να μειώνεται γραμμικά,

εκθετικά ή αντιστρόφως ανάλογα με το t (για παράδειγμα a(t)=0.9(1-t/100)).

Στο επόμενο στάδιο εκπαίδευσης, αποκαλούμενο και ως **tuning phase**, το εύρος γειτονιάς και ο ρυθμός μάθησης μπορούν είτε να μένουν σταθεροί σε μια πολύ μικρή τιμή είτε να μειώνονται πάρα πολύ αργά (έχοντας πάντως μια πολύ μικρή τιμή). Το στάδιο του tuning phase επιτελείται προκειμένου ο χάρτης να αποκτήσει μια τελική προσαρμογή (fine adjustment) των παραμέτρων του, τελειοποιώντας την περιγραφή της κατανομής των δεδομένων εισόδου. Ο ρυθμός μάθησης πρέπει να έχει μικρές τιμές για μια μακρά περίοδο (πχ μια τιμή της τάξεως του 0.02) χωρίς πάντως να είναι κρίσιμο το αν το a(t) θα φθίνει γραμμικά ή εκθετικά σε αυτή την τελική φάση.

Αντίθετα, στην περίπτωση ενός χάρτη μεγάλου μεγέθους, μπορεί να είναι σημαντικό να μειωθεί ο απαιτούμενος χρόνος εκπαίδευσης. Σε αυτή την περίπτωση, η επιλογή της συνάρτησης του ρυθμού μάθησης μπορεί να είναι αρκετά κρίσιμη για την επίδοση του δικτύου, με συνήθεις "βέλτιστες" επιλογές συναρτήσεις αντιστρόφως ανάλογες του t. Αποτελεσματικές επιλογές για τέτοιες συναρτήσεις και για τις παραμέτρους τους έχουν μέχρι στιγμής αποδειχθεί μόνο πειραματικά.

Επειδή ακριβώς η εκμάθηση του SOM αποτελεί μια στοχαστική διαδικασία, η τελική ακρίβεια της απεικόνισης εξαρτάται από τον αριθμό των βημάτων στην τελική φάση σύγκλισης (tuning phase), που πρέπει επομένως να είναι ικανοποιητικά μακρά. Αυτή η απαίτηση δεν γίνεται να αποφευχθεί. Μια εμπειρική τιμή για τη διάρκεια αυτής της φάσης είναι 500 φορές το μέγεθος του δικτύου. Επίσης, στην περίπτωση που μόνο λίγα δείγματα είναι διαθέσιμα, η διαδικασία επιτυγχάνεται μέσω της ανακύκλωσής τους.



Εικόνα 2.3. Δισδιάστατες κατανομές των δεδομένων εισόδου (περιοχές εντός πλαισίων) και τα δίκτυα των διανυσμάτων αναφοράς (νευρώνων) που τις προσεγγίζουν [1]

Δύο απλά παραδείγματα εκπαίδευσης

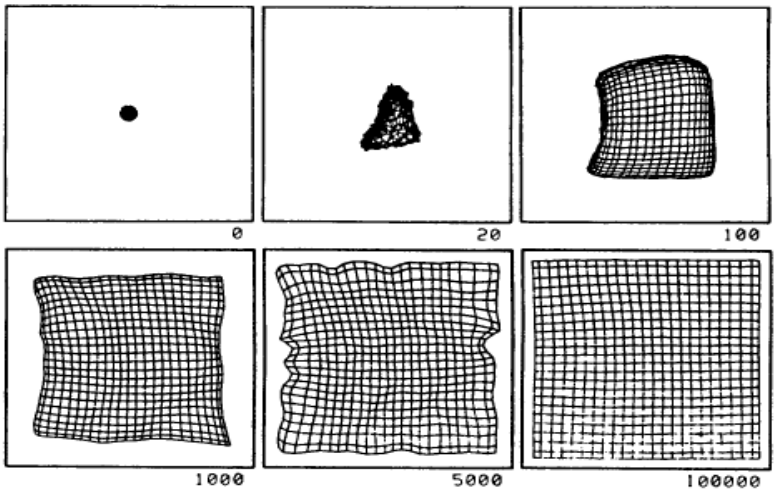
Στα παρακάτω απλά παραδείγματα, σκοπός μας είναι καταδειχθεί ότι ξεκινώντας από μια τυχαία αρχική κατάσταση των διανυσμάτων αναφοράς m_i, ο χάρτης μας διατάσσεται με ορθό τρόπο σε ένα πεπερασμένο αριθμό βημάτων. Σε πρακτικές εφαρμογές αντίθετα, προτιμάται η αρχική κατάσταση του χάρτη να είναι ήδη διατεταγμένη με τρόπο που να περιγράφει χονδρικά τη κατανομή των δεδομένων εισόδου, ώστε να επιτυγχάνεται γρήγορη σύγκλιση της διαδικασίας εκπαίδευσης, ακόμα και αν το εύρος της συνάρτησης γειτονιάς και η τιμή του ρυθμού εκμάθησης εκκινούν από αρκετά μικρές τιμές.

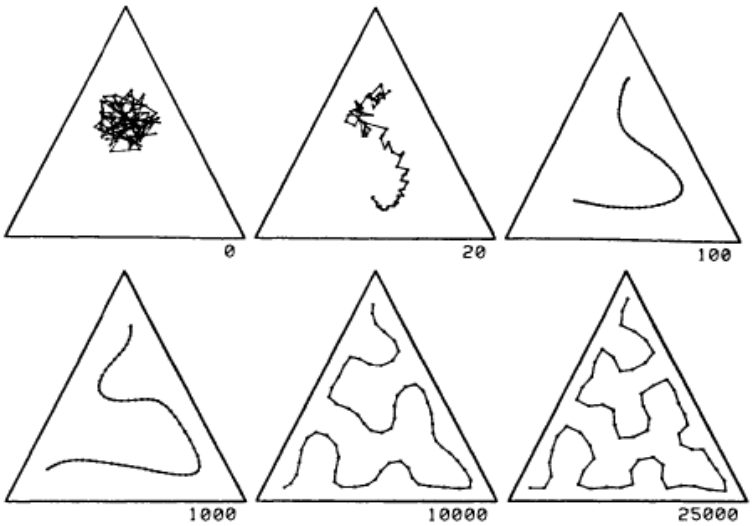
Θεωρούμε δεδομένα εισόδου δύο διαστάσεων $x = \left[\xi_1, \xi_2\right]^T \in \mathbb{R}^2$ με μια

αυθαίρετη δομή κατανομής. Για απλότητα έστω ότι το x είναι ένα στοχαστικό διάνυσμα, με τη συνάρτηση πυκνότητας πιθανότητας p(x) να θεωρείται ομοιόμορφη στο εσωτερικό των σχημάτων της εικόνας 2.3 και μηδενική οπουδήποτε αλλού. Στην εικόνα αυτή, οι τοπολογικές σχέσεις μεταξύ των νευρώνων σε ένα τετραγωνικό πίνακα, οπτικοποιούνται με βοηθητικές γραμμές και οι θέσεις των νευρώνων βρίσκονται πάνω στις διασταυρώσεις των γραμμών αυτών.

Τα διανύσματα αναφοράς κατά τη διάρκεια της διαδικασίας διάταξής τους, τείνουν να προσεγγίσουν την p(x) όχι με ακριβή πάντως τρόπο. Τα παραδείγματα της εικόνας 2.3 αναπαριστούν μια κατάσταση των διανυσματικών βαρών που έχει συγκλίνει προσεγγιστικά με τους διάφορους κόμβους του δικτύου να περιγράφουν διαφορετικές περιοχές των διανυσμάτων εισόδου, με διατεταγμένο τρόπο. Παρατηρούμε μια συστολή στα άκρα των δύο χαρτών, φαινόμενο που περιορίζεται με την αύξηση του μεγέθους του χάρτη.

Στις εικόνες 2.4α και 2.4β παρατηρούνται διάφορες ενδιάμεσες φάσεις της διαδικασίας αυτο-οργάνωσης του χάρτη. Σημειώνεται πως οι αρχικές τιμές των βαρών των κόμβων του δικτύου έχουν επιλεχθεί τυχαίες και έτσι η δομή του δικτύου γίνεται φανερή μετά από κάποιο χρονικό διάστημα εκπαίδευσης. Επίσης, το δίκτυο μπορεί να είναι διαφορετικής (μικρότερης) διάστασης σε σχέση με τα δεδομένα, παραδείγματος χάριν μονοδιάστατο όπως στην εικόνα 2.4β με την διάταξη να αποκτά μια μορφή καμπύλης Peano, ή ενός fractal.





Εικόνα 2.4. Διανύσματα αναφοράς κατά τη διάρκεια της ordering φάσης, α)τετραγωνικό δίκτυο β)γραμμικό (μίας διάστασης) δίκτυο . Οι αριθμοί κάτω δεξιά δείχνουν βήματα εκπαίδευσης. [1]

Αποτίμηση: Όταν ένας ικανοποιητικός αριθμός δειγμάτων x(t) έχει παρουσιαστεί και τα $m_i(t)$ έχουν συγκλίνει σε σχεδόν στάσιμες τιμές, το επόμενο βήμα είναι αποτίμηση (calibration) του χάρτη, προκειμένου να βρεθούν οι εικόνες των διαφορετικών δεδομένων εισόδου σε αυτόν. Στις πρακτικές εφαρμογές για τις οποίες τέτοιοι χάρτες χρησιμοποιούνται, μπορεί να είναι αυτονόητο πώς ένα ιδιαίτερο σύνολο δεδομένων εισόδου οφείλει να ερμηνευθεί και να ονομαστεί. Με την εισαγωγή διάφορων μη αυτόματα αναλυθέντων σύνολων δεδομένων και βλέποντας που επιτυγχάνονται οι καλύτερες αντιστοιχίες στο χάρτη σύμφωνα με την εξίσωση (2.1) και τοποθετώντας αντίστοιχες ετικέτες στους κόμβους του χάρτη, ο χάρτης αποτιμάται. Δεδομένου ότι αυτή η χαρτογράφηση υποτίθεται συνεχής κατά μήκος μιας υποθετικής «ελαστικής επιφάνειας», τα άγνωστα δεδομένα εισόδου προσεγγίζονται από τα πιο κοντινά διανύσματα αναφοράς, όπως στον κβαντισμό διανυσμάτων.

Σχόλιο: Μια «βέλτιστη χαρτογράφηση» μπορεί να θεωρηθεί ως μια που προβάλλει την συνάρτηση πυκνότητας πιθανότητας p(x) με τον πιο "πιστό" τρόπο, προσπαθώντας τουλάχιστον να διατηρήσει τις τοπικές δομές της p(x) στο χώρο εξόδου.

Πρέπει να υπογραμμιστεί, εντούτοις, ότι η περιγραφή της ακριβούς μορφής της p(x) από το SOM δεν είναι ο σημαντικότερος στόχος. Θα γίνει φανερό στη συνέχεια, ότι το SOM βρίσκει αυτόματα εκείνες τις διαστάσεις και περιοχές στο χώρο εισόδου όπου το x έχει μεγαλύτερο αριθμό δειγμάτων, υπακούοντας έτσι στη συνηθισμένη φιλοσοφία των προβλημάτων οπισθοδρόμησης.

2.4 SOM με μετρική εσωτερικού γινομένου

Σε κάποιες περιπτώσεις έχει προταθεί το x να είναι κανονικοποιημένο προτού να χρησιμοποιηθεί στον αλγόριθμο. Αν και η κανονικοποίηση δεν είναι απαραίτητη σε γενικές γραμμές, μπορεί να βελτιώσει την αριθμητική ακρίβεια επειδή τα προκύπτοντα από αυτήν διανύσματα αναφοράς τείνουν να κινούνται στην ίδια δυναμική περιοχή.

Μια άλλη πτυχή είναι ότι είναι δυνατό να εφαρμοστούν πολλές διαφορετικές μετρικές για την αντιστοίχηση, εντούτοις, οι κανόνες αντιστοίχησης και ενημέρωσης οφείλουν να είναι αμοιβαία συμβατοί όσον αφορά την ίδια μετρική. Για παράδειγμα, εάν εφαρμοστεί το εσωτερικό γινόμενο για να ορίσει την ομοιότητα των x και mi, οι εξισώσεις εκμάθησης αποκτούν τη μορφή

$$\boldsymbol{x}^T(t)\boldsymbol{m}_c(t)=\max_i\left\{\boldsymbol{x}^T(t)\boldsymbol{m}_i(t)\right\}, \tag{2.5}$$

$$\boldsymbol{m}_i(t+1)=\begin{cases}\dfrac{\boldsymbol{m}_i(t)+a'(t)\boldsymbol{x}(t)}{\|\boldsymbol{m}_i(t)+a'(t)\boldsymbol{x}(t)\|} & \text{if } i\in N_c(t)\\ \boldsymbol{m}_i(t) & \text{if } i\notin N_c(t)\end{cases} \tag{2.6}$$

και $0 < a'(t) < \infty$ για παράδειγμα, $a'(t) = 100/t$. Αυτή η διαδικασία κανονικοποιεί αυτόματα τα διανύσματα αναφοράς σε κάθε βήμα. Οι υπολογισμοί κανονικοποίησης βέβαια, επιβραδύνουν τον αλγόριθμο εκπαίδευσης. Από την άλλη πλευρά, κατά τη διάρκεια της αντιστοίχησης, το κριτήριο εσωτερικού γινομένου που εφαρμόζεται για την αναγνώριση των προτύπων είναι πολύ απλό και γρήγορο, και υποκείμενο σε πολλά είδη απλού αναλογικού υπολογισμού, ηλεκτρονικού και οπτικού.

Εκτός από τις ευκλείδειες αποστάσεις και το εσωτερικό γινόμενο, πολλά άλλα κριτήρια αντιστοίχησης μπορούν να χρησιμοποιηθούν με το SOM (ένα τέτοιο, διαφορετικό κριτήριο αντιστοίχησης χρησιμοποιείται άλλωστε και στον ορισμό του υβριδικού SOHMMM όπως θα δούμε αργότερα).

2.5 Αρχικοποίηση του αλγόριθμου του SOM

Τυχαία αρχικοποίηση. Ο λόγος για τον οποίο χρησιμοποιήθηκαν τυχαίες αρχικές τιμές στις επιδείξεις του SOM ήταν ότι το SOM μπορεί να αρχικοποιηθεί χρησιμοποιώντας αυθαίρετες τιμές για τα διανύσματα αναφοράς mi(0). Με άλλα λόγια έχει καταδειχθεί ότι τα αρχικά μη ταξινομημένα διανύσματα θα διαταχτούν μακροπρόθεσμα, στις συνηθισμένες εφαρμογές, σε μερικές εκατοντάδες βήματα. Αυτό δεν σημαίνει, εντούτοις, ότι η τυχαία αρχικοποίηση είναι η καλύτερη ή γρηγορότερη πολιτική και πρέπει να χρησιμοποιείται στην πράξη.

Γραμμική αρχικοποίηση. Αφού τα mi(0) μπορεί να είναι αυθαίρετα, θα μπορούσαμε να ισχυριστούμε ότι οποιαδήποτε διατεταγμένη αρχική κατάσταση είναι επιθυμητή, ακόμα κι αν οι αρχικές τιμές δεν βρίσκονται κατά μήκος των κύριων επεκτάσεων της σππ p(x). Μια μέθοδος που έχει χρησιμοποιηθεί επιτυχώς είναι αρχικά να καθοριστούν δύο ιδιοδιανύσματα της μήτρας αυτοσυσχέτισης του x που έχουν τις μεγαλύτερες ιδιοτιμές, και στη συνέχεια να εκταθούν δημιουργώντας ένα δισδιάστατο γραμμικό υποχώρο. Ένας ορθογωνικός πίνακας

(με ορθογωνικό ή εξαγωνικό κανονικό πλέγμα) καθορίζεται κατά μήκος αυτού του υποχώρου, με το κέντρο του να συμπίπτει με αυτό του μέσου όρου του x(t), και τις κύριες διαστάσεις να είναι ίδιες με τις δύο μεγαλύτερες ιδιοτιμές. Οι αρχικές τιμές mi(0) προσδιορίζονται έπειτα από τα σημεία του πίνακα. Εάν κάποιος θέλει να λάβει ένα προσεγγιστικά ομοιόμορφο πλέγμα στο SOM, οι σχετικές τιμές των κόμβων στις οριζόντιες και κάθετες κατευθύνσεις του πλέγματος, αντίστοιχα, πρέπει να είναι ανάλογες προς τις δύο μεγαλύτερες ιδιοτιμές που προσδιορίσθηκαν προηγουμένως.

Αφού τα mi(0) έχουν ήδη διαταχθεί και η πυκνότητα σημείου τους προσεγγίζει χονδρικά την p(x), θα είναι δυνατό να εκκινήσει άμεσα η εκμάθηση με τη φάση σύγκλισης, όπου μπορούν να χρησιμοποιηθούν τιμές για το ρυθμό εκμάθησης α(t) σημαντικά μικρότερες από την μονάδα, και μια συνάρτηση γειτονιάς το πλάτος της οποίας είναι κοντά στην τελική τιμή της, πλησιάζοντας την ισορροπία ομαλά.

2.6 Βέλτιστος ρυθμός εκμάθησης

Η χρονική περίοδος κατά τη διάρκεια της οποίας επιτυγχάνεται μια χονδρική διάταξή στο SOM είναι συνήθως σχετικά σύντομη, της τάξεως των 1000 βημάτων, ενώ το μεγαλύτερο μέρος του χρόνου υπολογισμού αφιερώνεται για την τελική φάση σύγκλισης, προκειμένου να επιτευχθεί μια αρκετά καλή στατιστική ακρίβεια. Δεν είναι σαφές πώς ο παράγοντας εκμάθησης θα πρέπει να βελτιστοποιηθεί κατά τη διάρκεια της πρώτης φάσης, αφού το πλάτος της συνάρτησης γειτονιάς αλλάζει με αυτό τον τρόπο επίσης, κάτι που περιπλέκει την κατάσταση.

Από την άλλη πλευρά, στη προηγούμενη παράγραφο ήδη επισημάνθηκε ότι είναι πάντα δυνατό να αρχίσει ο αλγόριθμος του SOM με μια ήδη διατεταγμένη κατάσταση, πχ. με όλα το mi σε μια κανονική διάταξη, κατά μήκος μιας δισδιάστατης επιφάνειας. Επίσης, δεδομένου ότι κατά τη διάρκεια της τελικής φάσης σύγκλισης κρατάμε συνήθως το πλάτος της γειτονιάς σταθερό, φαίνεται δυνατό να καθοριστεί κάποιο είδος «βέλτιστου» νόμου για την ακολουθία {α(t)} κατά τη διάρκεια της φάσης σύγκλισης. Προτού παρουσιαστεί αυτός ο νόμος για το SOM θα συζητηθούν μερικά απλούστερα παραδείγματα, από τα οποία η βασική ιδέα μπορεί να γίνει αντιληπτή.

Αναδρομική μέθοδος. Θεωρούμε ένα σύνολο διανυσματικών δειγμάτων {x(t)}διάστασης n με t = 0.1.2,…, T. Από τον τρόπο με τον οποίο υπολογίζεται αναδρομικά ο μέσος m των x(t) μπορεί γίνει σαφές ότι το m(t) είναι πάντα σωστό όσον αφορά τα δείγματα που λαμβάνονται υπόψη μέχρι το βήμα t + 1:

$$m = \frac{1}{T}\sum_{t=1}^{T} x(t) = m(T) \qquad (2.7)$$

$$m(t+1) = \frac{t}{t+1}m(t)+\frac{1}{t+1}x(t+1) = m(t)+\frac{1}{t+1}\left[x(t+1)-m(t)\right] \qquad (2.8)$$

Αυτός ο νόμος θα απεικονιστεί τώρα και στις διαδικασίες κβαντισμού διανυσμάτων.

«Βελτιστοποιημένος» ρυθμός μάθησης για τον αναδρομικό VQ. Η αναδρομή της μεθόδου απότομης κατάβασης (steepest descent) για τον κλασικό κβαντισμό διανυσμάτων έχει τη μορφή

mi(t+1) = mi(t) + α (t)δci [x(t) -mi(t)] **(2.9)**

όπου δci είναι η συνάρτηση δέλτα του Kronecker. Η (2.9) μπορεί να ξαναγραφεί ως
mi(t+1) = [1-α (t)δci]mi(t)+α (t)δci x(t) **(2.10)**

Παρατηρούμε μια μετατόπιση στο δείκτη x(t) σε σχέση με την (2.8). Είναι θέμα σύμβασης το πώς ονομάζουμε το x(t). Εάν θεωρήσουμε το mi (t+1) σαν μια μνήμη όλων των τιμών x(t'), t'=0,1,...t βλέπουμε ότι αν το mi είναι ο κόμβος νικητής, τότε το x(t) αφήνει στην τιμή του mi(t) ένα "ίχνος", σταθμισμένο με τον ρυθμό μάθησης α(t). Αν το mi ήταν «νικητής» στο βήμα t, θα έχει ένα ίχνος μνήμης του x(t-1), σταθμισμένο με [1 - α(t)]α(t - 1) μέσω του πρώτου όρου της 2.10. Προφανώς, εάν το mi(t) δεν ήταν νικητής, κανένα ίχνος μνήμης από το x(t-1). δεν λαμβάνεται. Είναι τώρα εύκολο να δει κανείς τι συμβαίνει στα προηγούμενα ίχνη μνήμης x(t'): κάθε φορά που το mi είναι νικητής, όλα το x(t') μειώνονται κατά τον παράγοντα [1 - α(t)], ο οποίος υποτίθεται < 1. Εάν το α(t) ήταν σταθερό στο χρόνο, μακροπρόθεσμα η επίδραση του x(t') στο m(t+1) θα *ξεχαστεί*, δηλαδή το m(t+1) εξαρτάται μόνο από τα σχετικά λίγα τελευταία δείγματα, με τον «χρόνο λήθης» να εξαρτάται από την τιμή του α. Εάν, αντιθέτως, οι αρχικές τιμές του α(t') επιλέγονται να είναι μεγαλύτερες, είναι πιθανότερο να είμαστε σε θέση να αντισταθμίσουμε την «λήθη» και να έχουμε μια κατά προσέγγιση όμοια επιρροή του x(t') στο m(t + 1).

Σχόλιο. Δεδομένου ότι το διάγραμμα Voronoi στον κβαντισμό διανυσμάτων (VQ) αλλάζει σε κάθε βήμα, τα συστατικά από τα οποία οι τιμές των x(t) προέρχονται για ένα συγκεκριμένο mi(t) αλλάζουν επίσης, έτσι αυτή η περίπτωση δεν είναι τόσο σαφής όσο για την περίπτωση της απλής αναδρομικής μεθόδου. Επομένως οι τιμές που προσδιορίζουμε για τον ρυθμό μάθησης α(t) μπορούν να είναι μόνο προσεγγιστικά βέλτιστες.

Θεωρούμε τώρα τις δύο πιο κοντινές στιγμές t1 και t2, με t1 < t2, για τις οποίες ο ίδιος κόμβος mi, συμβολιζόμενος ως mc, ήταν ο κόμβος νικητής. Εάν ορίσουμε ότι τα ίχνη μνήμης των x(t1) και x(t2) θα είναι ίσα σε διαφορετικούς νευρώνες σε όλες τις περιπτώσεις, προφανώς πρέπει να επιλέξουμε έναν μεμονωμένο ρυθμό μάθησης ai(t) για κάθε mi. Για τα βήματα t1 και t2 πρέπει σε αυτή την περίπτωση να έχουμε
[1 – αc(t2)]αc(t1) = αc(t2) **(2.11)**
Επιλύοντας ως προς αc(t2) από την (2.11) παίρνουμε

$$\alpha_c(t_2) = \frac{\alpha_c(t_1)}{1+\alpha_c(t_1)} \qquad \textbf{(2.12)}$$

Υπογραμμίζεται ότι το αi(t) τώρα αλλάζει μόνο όταν το mi ήταν ο «νικητής», διαφορετικά το αi(t) παραμένει αμετάβλητο. Δεδομένου ότι το αc(t1) διατηρεί την ίδια τιμή μέχρι το βήμα t2 -1, μπορούμε εύκολα να δούμε ότι η ακόλουθη ανάδρομη μπορεί να εφαρμοστεί συγχρόνως όταν ανανεώνεται ο κόμβος νικητής mc:

$$\alpha_c(t+1) = \frac{\alpha_c(t)}{1+\alpha_c(t)} \qquad \textbf{(2.13)}$$

Πρέπει άλλη μια φορά να υπογραμμιστεί ότι η (2.13), καθώς επίσης και οι παρόμοιες εκφράσεις που παράγονται για το SOM και την LVQ, δεν εγγυώνται την απολύτως βέλτιστη σύγκλιση, επειδή τα διαγράμματα Voronoi σε αυτούς τους αλγορίθμους αλλάζουν. Εντούτοις, μπορούμε να πούμε με βεβαιότητα ότι εάν η (2.13) δεν λαμβάνεται υπόψη, η σύγκλιση είναι λιγότερο βέλτιστη κατά μέσο όρο.

«Βελτιστοποιημένος» ρυθμός μάθησης για το SOM. Είναι τώρα άμεσα δυνατό να γενικευθεί η ανωτέρω φιλοσοφία για το SOM. Εάν ορίσουμε ένα μεμονωμένο αi(t) για κάθε

mi και γράψουμε

m(t +1) = mi(t)+αi(t)hci[x(t)-mi(t)] **(2.14)**

όπου η συνάρτηση γειτονιάς hci (στη φάση σύγκλισης) είναι χρονικά αμετάβλητη, μπορούμε να ενημερώνουμε το αi(t) όποτε πραγματοποιείται μια διόρθωση στις τιμές του mi:

$$\alpha_i(t+1) = \frac{\alpha_i(t)}{1+h_{ci}\alpha_i(t)} \quad (2.15)$$

Ο παραπάνω νόμος για το αi(t) βασίστηκε σε μια θεωρητική εικασία και είναι πιθανό να μην λειτουργήσει με τον καλύτερο τρόπο στην πράξη, λόγω των πολύ διαφορετικών τιμών για τα αi(t) που λαμβάνονται μακροπρόθεσμα για τα διαφορετικά i. Υπενθυμίζεται ότι η πυκνότητα σημείου των διανυσμάτων αναφοράς στον αρχικό αλγόριθμο SOM ήταν κάποια μονοτονική συνάρτηση της p(x). Αυτή η πυκνότητα επηρεάζεται από τις διαφορές στους ρυθμούς εκμάθησης.

Παρολ' αυτά, στον κβαντισμό διανυσμάτων και ειδικά στον LVQ, αυτή η ιδέα λειτουργεί εύλογα καλά και δεν παρατηρείται κάποια επιβλαβής παραμόρφωση των πυκνοτήτων σημείου των mi.

Ημι-εμπειρικός ρυθμός εκμάθησης. Για τους προαναφερθέντες λόγους φαίνεται δικαιολογημένο να κρατηθούν οι α(t) στο SOM ίδιοι για όλους τους νευρώνες και να αναζητηθεί ένας μέσος βέλτιστος ρυθμός. Οι Mulier και Cherkassky [5] κατέληξαν σε μια έκφραση της μορφής

$$\alpha(t) = \frac{A}{t+B} \quad (2.16)$$

όπου τα A και B είναι κατάλληλα επιλεγμένες σταθερές. Αυτή η μορφή ικανοποιεί τουλάχιστον τις στοχαστικές συνθήκες. Η κύρια αιτιολόγηση για την εξίσωση (2.16) είναι ότι τα προηγούμενα και πιο πρόσφατα δείγματα θα ληφθούν υπόψη με όμοια περίπου μέσα βάρη.

2.7 Επίδραση της μορφής της συνάρτησης γειτονιάς

Εφ' όσον κάποιος αρχίζει τη διαδικασία αυτο-οργάνωσης με μια ευρεία συνάρτηση γειτονιάς, δηλαδή με μια ευρεία ακτίνα του συνόλου γειτονιάς Nc(0), ή μια ευρεία τυπική απόκλιση hci(0) με μια τιμή που να είναι τάξης μεγέθους της μισής από τη μεγαλύτερη διάσταση του πίνακα, δεν υπάρχει συνήθως κανένας κίνδυνος να καταλήξουμε σε «μετα-σταθερές» μορφές του χάρτη (για τις οποίες η μέση αναμενόμενη παραμόρφωση (distortion measure)ή το μέσο αναμενόμενο σφάλμα κβαντισμού θα κατέληγαν σε τοπικό και όχι σε ολικό ελάχιστο). Εντούτοις, με την χρονικά αμετάβλητη συνάρτηση γειτονιάς η κατάσταση μπορεί να είναι αρκετά διαφορετική, ειδικά εάν η συνάρτηση γειτονιάς είναι στενή.

Ο Erwin et al.[6] έχει αναλύσει τις «ασταθείς καταστάσεις» σε έναν μονοδιάστατο πίνακα. Καθόρισαν αρχικά τη λειτουργία γειτονιάς να είναι κυρτή σε ένα ορισμένο διάστημα I = {0.1.2,…, N}, εάν οι συνθήκες |s - q| >|s - r| και |s - q|>|r – q| υπονοούν ότι [h(s, s) + h(s, q)] < [h(s, r) + h(r, q)] για όλα τα s, r, q που ανήκουν στο I. Διαφορετικά η λειτουργία γειτονιάς θεωρήθηκε κοίλη.

Τα κύρια αποτελέσματα στα οποία κατέληξαν ήταν ότι εάν η συνάρτηση γειτονιάς είναι κυρτή, δεν υπάρχει καμία άλλη σταθερή κατάσταση εκτός από αυτή της διατεταγμένης. Εάν η συνάρτηση γειτονιάς είναι κοίλη, υπάρχουν ασταθείς καταστάσεις που μπορούν να

επιβραδύνουν τη διαδικασία διάταξης κατά τάξεις μεγέθους. Επομένως, εάν στην αρχή της διαδικασίας η συνάρτηση γειτονιάς είναι κυρτή, όπως στο μεσαίο τμήμα της γκαουσσιανής hci(t) με μεγάλη τυπική απόκλιση, η διάταξη του χάρτη μπορεί να επιτευχθεί σχεδόν σίγουρα. Στη συνέχεια η συνάρτηση γειτονιάς μπορεί να συρρικνωθεί (γίνει στενότερη) για να επιτευχθεί μια βελτιωμένη προσέγγιση της p(x).

Οι συνθήκες διάταξης είναι γενικά αυστηρότερες εάν ο χώρος των σημάτων εισόδου έχει την ίδια διάσταση με τον πίνακα, στην πράξη, εντούτοις, η διάσταση του χώρου των σημάτων εισόδου είναι συνήθως πολύ υψηλότερη, οπότε η διάταξη του SOM πραγματοποιείται ευκολότερα.

2.8 Ένα κριτήριο παραμόρφωσης για τον αλγόριθμο του SOM

Δεδομένου ότι το SOM ανήκει στην κατηγορία μεθόδων κβαντισμού διανυσμάτων (VQ), μπορεί να υποτεθεί ότι αφετηρία στη βελτιστοποίηση του πρέπει να αποτελεί κάποιο είδος σφάλματος κβαντισμού(quantization error) στο διανυσματικό χώρο. Έστω ότι το $x \in \mathbb{R}^n$ είναι το διάνυσμα εισόδου και το $m_i \in \mathbb{R}^n$ (με i δείκτη νευρώνων) είναι τα διανύσματα αναφοράς και έστω ότι το d(x,mi) καθορίζει μια γενικευμένη συνάρτηση απόστασης μεταξύ των x και mi. Το σφάλμα κβαντισμού θα ορίζεται ως

$$d(x, m_c) = \min_i \{d(x, m_i)\} \tag{2.17}$$

όπου με c συμβολίζουμε το «πιο κοντινό» στο x διάνυσμα αναφοράς στον χώρο των σημάτων εισόδου.

Μια σημαντικότερη συνάρτηση στο SOM, εντούτοις, είναι η συνάρτηση γειτονιάς hci= hci(t) που περιγράφει την αλληλεπίδραση των διανυσμάτων αναφοράς mi και mc κατά τη διάρκεια της εκπαίδευσης και είναι συνήθως μια συνάρτηση του χρόνου t. Με βάση την παραπάνω παρατήρηση μπορεί να οριστεί ένα μέτρο παραμόρφωσης με τον ακόλουθο τρόπο:

Συμβολίζοντας το σύνολο των δεικτών όλων των κόμβων του πλέγματος με L το μέτρο παραμόρφωσης e ορίζεται ως

$$e = \sum_{i \in L} h_{ci} d(x, m_i) \tag{2.18}$$

δηλαδή ως το άθροισμα των συναρτήσεων απόστασης σταθμισμένων από την hci, όπου c είναι ο δείκτης του πιο κοντινού στο x διανύσματος αναφοράς. Στη συνέχεια μπορούμε να ορίσουμε το μέσο αναμενόμενο μέτρο παραμόρφωσης:

$$E = \int ep(x)dx = \int \sum_{i \in L} h_{ci} d(x, m_i) p(x) dx \tag{2.19}$$

ένας τρόπος ορισμού του SOM είναι ως το σύνολο των mi που ελαχιστοποιεί (ολικά) το E. Η ακριβής βελτιστοποίηση της εξίσωσης (2.19), εντούτοις, είναι προς το παρόν ένα άλυτο θεωρητικό πρόβλημα και είναι εξαιρετικά υψηλής υπολογιστικής πολυπλοκότητας. Η καλύτερη προσεγγιστική λύση που έχει βρεθεί μέχρι σήμερα είναι βασισμένη στη στοχαστική προσέγγιση Robbins-Monro. Σύμφωνα με αυτήν την ιδέα εξετάζουμε τα στοχαστικά δείγματα του μέτρου παραμόρφωσης: εάν {x(t), t=1, 2….} είναι μια ακολουθία δειγμάτων και {mi(t), t=1, 2....} η αναδρομικά ορισμένη ακολουθία του διανύσματος αναφοράς mi τότε

$$e = \sum_{i \in L} h_{ci}(t) d(x(t), m_i(t)) \tag{2.20}$$

είναι μια στοχαστική μεταβλητή, και η ακολουθία που καθορίζεται από τα

$$m_i(t+1) = m_i(t) - \lambda \cdot \nabla_{m_i} e(t) \tag{2.21}$$

χρησιμοποιείται για να βρεθεί μια προσέγγιση στη βέλτιστη λύση, με τη μορφή ασυμπτωτικών τιμών του mi. Αυτό θα καθόριζε συνεπώς τον αλγόριθμο του SOM για μια γενικευμένη συνάρτηση απόστασης d(x, mi). Πρέπει να υπογραμμιστεί, εντούτοις, ότι αν και οι ιδιότητες σύγκλισης της στοχαστικής προσέγγισης ήταν γνωστές λεπτομερώς από το 1951, οι ασυμπτωτικές τιμές του mi που λαμβάνονται από την (2.21) ελαχιστοποιούν την E προσεγγιστικά. Σε κάθε περίπτωση οι εξισώσεις (2.20) και (2.21) μπορούν να θεωρηθούν ως ένας ακόμα ορισμός μιας κλάσης αλγορίθμων SOM.

Σχόλιο Σε αυτό το σημείο θα ήταν καλό να υπενθυμιστεί ότι το E δεν χρειάζεται να είναι μια συνάρτηση ορμής και το όριο σύγκλισης δεν αντιπροσωπεύει απαραιτήτως το ακριβές ελάχιστο του E, μόνο μια προσέγγιση του. Εντούτοις, οι ιδιότητες σύγκλισης αυτής της κατηγορίας διαδικασιών είναι πολύ ισχυρές και έχουν μελετηθεί λεπτομερώς πολύ καιρό πριν σε θεωρητικό επίπεδο. Επίσης, Προηγουμένως, η συνάρτηση γειτονιάς hci υποτέθηκε χρονικά αμετάβλητη. Φαίνεται πάντως, να είναι μια σημαντική ιδιότητα του αλγόριθμου του SOM η hci να μεταβάλλεται στο χρόνο.

2.9 Υποδείξεις για την κατασκευή καλών χαρτών (SOM)

Αν και είναι δυνατό να ληφθούν κάποια είδη χαρτών χωρίς κάποια ιδιαίτερη προσοχή στην επιλογή παραμέτρων, εν τούτοις, είναι χρήσιμο να δοθεί προσοχή στις ακόλουθες παρατηρήσεις ώστε οι προκύπτουσες απεικονίσεις να είναι σταθερές, καλά προσανατολισμένες και όσο το δυνατόν ελάχιστα διφορούμενες. Το πρόβλημα της αρχικοποίησης του SOM έχει ήδη συζητηθεί στην παράγραφο 2.5.

Μορφή του πλέγματος. Για οπτικοποίηση, προτιμάται το εξαγωνικό πλέγμα, επειδή δεν ευνοεί τις οριζόντιες και κάθετες κατευθύνσεις τόσο όσο το ορθογωνικό. Τα άκρα του πλέγματος οφείλουν να είναι ορθογωνικά αντί τετραγωνικά, επειδή το «ελαστικό δίκτυο» που σχηματίζεται από τα διανύσματα αναφοράς mi πρέπει να προσανατολιστεί κατά τη μορφή της p(x) και να σταθεροποιηθεί κατά τη διαδικασία εκμάθησης. Παρατηρήστε ότι εάν το πλέγμα ήταν, πχ. κυκλικό, δεν θα είχε κανέναν σταθερό προσανατολισμό στο χώρο δεδομένων. Συνεπώς, οποιαδήποτε επιμηκυμένη μορφή είναι προτιμότερη. Παρολ' αυτά, δεδομένου ότι τα mi πρέπει να προσεγγίσουν την συνάρτηση πυκνότητας πιθανότητας p(x), θα ήταν επιθυμητό ο χάρτης να έχει τέτοιες διαστάσεις που να αντιστοιχούν κατά προσέγγιση στις σημαντικότερες διαστάσεις της p(x). Επομένως, θα ήταν καλό να προηγείται μια εποπτεία, χονδρικά, της μορφής της p(x), πχ., με το Sammon's mapping [7]

Εκπαίδευση με έναν μικρό αριθμό διαθέσιμων δειγμάτων. Δεδομένου ότι για να έχουμε μια καλή στατιστική ακρίβεια, η πλήρης διαδικασία εκμάθησης μπορεί να απαιτήσει έναν αριθμό, ενδεικτικά, περίπου 100.000 βημάτων και ο αριθμός διαθέσιμων δειγμάτων είναι συνήθως πολύ μικρότερος, είναι προφανές ότι τα δείγματα πρέπει να χρησιμοποιηθούν επαναληπτικά στην εκπαίδευση. Διάφορες εναλλακτικές λύσεις υπάρχουν προς τούτο: τα δείγματα μπορούν να εφαρμοστούν κυκλικά ή σε μια τυχαία μεταβαλλόμενη διάταξή, ή να επιλεγούν τυχαία από το βασικό σύνολο (bootstrap learning). Έχει αποδειχθεί στην πράξη ότι η κυκλική εφαρμογή των δειγμάτων δεν είναι καταφανώς χειρότερη από τις άλλες, από μαθηματική απόψεως καλύτερα δικαιολογήσιμες, μεθόδους.

Βελτίωση των σπάνιων περιπτώσεων. Μπορεί επίσης να είναι προφανές από τα ανωτέρω ότι το SOM τείνει με τον ένα ή τον άλλο τρόπο να αντιπροσωπεύσει την p(x). Εντούτοις, σε

πολλά πρακτικά προβλήματα σημαντικές περιπτώσεις (δεδομένων εισόδου) μπορεί να εμφανιστούν με τη μικρή στατιστική συχνότητα, οπότε δεν είναι ικανά να καταλάβουν κάποιο τμήμα στο χώρο απεικόνισης (SOM). Επομένως, τέτοιες σημαντικές περιπτώσεις μπορούν να ενισχυθούν στην εκμάθηση αυθαίρετα, με τη χρήση μιας υψηλότερης τιμής για το α ή το hci για αυτά τα δείγματα, ή την επανάληψη αυτών των δειγμάτων σε μια τυχαία διάταξη για έναν ικανοποιητικό αριθμό φορών κατά τη διάρκεια της διαδικασίας εκμάθησης. Είναι φανερό πως ο προσδιορισμός της ζητούμενης ενίσχυσης κάποιων δειγμάτων κατά την εκμάθηση πρέπει να γίνεται σε συνεργασία με τους τελικούς χρήστες αυτών των χαρτών.

Κανονικοποίηση των συστατικών μοτίβων. Αυτό είναι ένα πολύ λεπτό πρόβλημα. Γίνεται εύκολα αντιληπτό ότι ο προσανατολισμός, ή η διατεταγμένη «οπισθοδρόμηση» των διανυσμάτων αναφοράς στο χώρο εισόδου πρέπει να εξαρτάται από την κανονικοποίηση των συστατικών (ή των διαστάσεων) των διανυσμάτων εισόδου. Εντούτοις, εάν τα δεδομένα εισόδου έχουν αντιπροσωπευθεί ήδη σε διαφορετικές κλίμακες, δεν υπάρχει κάποιος απλός κανόνας για να καθορίσει τι είδους βέλτιστη κανονικοποίηση πρέπει να χρησιμοποιηθεί πριν εισαχθούν τα δεδομένα εκπαίδευσης στον αλγόριθμο εκμάθησης. Η ευκολότερη λύση, που είναι συνήθως μάλλον αποτελεσματική, ειδικά με δεδομένα εισόδου υψηλής διάστασης, είναι να κανονικοποιηθεί η διακύμανση κάθε συστατικού επί των δεδομένων εκπαίδευσης. Είναι επίσης εφικτό να δοκιμαστούν πολλές ευριστικά δικαιολογήσιμες κανονικοποιήσεις και να ελεγχθεί η ποιότητα των προκυπτόντων χαρτών με τη βοήθεια του Sammon mapping ή των μέσων σφαλμάτων κβαντισμού.

Επιβάλλοντας απεικονίσεις δεδομένων σε μια επιθυμητή θέση στο χάρτη. Μερικές φορές, ειδικά όταν το SOM χρησιμοποιείται για να ελέγξει τα πειραματικά δεδομένα, μπορεί να είναι επιθυμητό να χαρτογραφηθούν τα «κανονικά» δεδομένα σε μια καθορισμένη θέση (πχ, στη μέση) του χάρτη. Προκειμένου να "αναγκαστούν" να τοποθετηθούν τα συγκεκριμένα στοιχεία στις επιθυμητές θέσεις, είναι ενδεδειγμένο να χρησιμοποιηθούν τα αντίγραφά τους για τις αρχικές τιμές των διανυσμάτων αναφοράς σε αυτές τις θέσεις, και να κρατηθεί ο ρυθμός εκμάθησης α χαμηλός για αυτές τις θέσεις-κόμβους κατά τη διάρκεια της ενημέρωσής τους.

Έλεγχος της ποιότητας της εκμάθησης. Οι διαφορετικές διαδικασίες εκμάθησης μπορούν να οριστούν ξεκινώντας με διαφορετικές αρχικές τιμές mi(0), και εφαρμόζοντας διαφορετικές ακολουθίες διανυσμάτων εκπαίδευσης x(t) και διαφορετικές παραμέτρους εκμάθησης. Είναι προφανές ότι μπορεί να υπάρξει κάποιος βέλτιστος χάρτης για τα ίδια δεδομένα εισόδου. Μπορεί επίσης να είναι προφανές ότι ο καλύτερος χάρτης αναμένεται να επιδεικνύει το μικρότερο μέσο σφάλμα κβαντισμού, τουλάχιστον προσεγγιστικά, επειδή ταιριάζει έπειτα καλύτερα στα ίδια δεδομένα. Ο μέσος όρος του ||x-mc||, που καθορίζεται μέσω της εισαγωγής των δεδομένων εκπαίδευσης μια επιπλέον φορά μετά το πέρας της εκπαίδευσης, είναι ένας χρήσιμος δείκτης απόδοσης. Επομένως, οφείλουν να δοκιμαστούν ένας αξιόλογος αριθμός (αρκετές δεκάδες) τυχαίων αρχικοποιήσεων των mi(0) και διαφορετικές ακολουθίες εκμάθησης, και να επιλεχτεί ο χάρτης με το ελάχιστο σφάλμα κβαντισμού.

Η ακρίβεια των χαρτών στη διατήρηση της τοπολογίας, ή οι σχέσεις γειτονικών κόμβων, του χώρου εισόδου έχουν μετρηθεί με διάφορους τρόπους. Μια προσέγγιση είναι να συγκριθούν οι σχετικές θέσεις των διανυσμάτων αναφοράς (στο χώρο εισόδου) με τις σχετικές θέσεις των αντίστοιχων κόμβων στο χάρτη [8]. Για παράδειγμα, μπορεί να μετρηθεί ο αριθμός φορών που η περιοχή Voronoi ενός άλλου κόμβου χάρτη «παρεισφρέει» στη μέση των διανυσμάτων αναφοράς δύο κόμβων γειτόνων [9]. Μια διαφορετική προσέγγιση είναι να εξεταστεί, για κάθε διάνυσμα εισόδου, η απόσταση της καλύτερα-ταιριάζουσας μονάδας

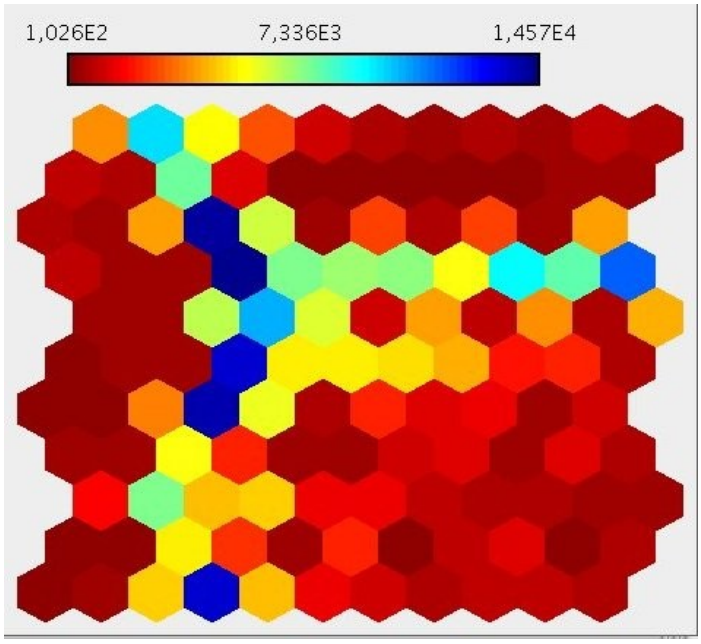
(BMU) από τη δεύτερη-καλύτερα-ταιριάζουσα μονάδα στο χάρτη: Εάν οι μονάδες δεν είναι γειτονικές, η τοπολογία δεν διατηρείται ([10], [11]).

Όταν η απόσταση μεταξύ των μονάδων χαρτών καθορίζεται κατάλληλα, ένα τέτοιο μέτρο μπορεί να συνδυαστεί με το σφάλμα κβαντισμού για να διαμορφώσει ένα ενιαίο μέτρο της ποιότητας ενός χάρτη [12]. Αν και δεν είναι αυτή τη στιγμή δυνατό να υποδειχθεί το καλύτερο μέτρο της ποιότητας χαρτών, αυτά τα μέτρα μπορούν εντούτοις να είναι χρήσιμα στην επιλογή των κατάλληλων παραμέτρων εκμάθησης και των μεγεθών των χαρτών.

2.10 Οπτικοποίηση για την επίδειξη των συστάδων στο SOM

Ένα πρόβλημα που θέλουμε να επιλύσουμε μετά το πέρας της εκπαίδευσης του SOM, είναι να υπάρχει μια όσο το δυνατόν πιο ξεκάθαρη οπτικοποίηση των διάφορων συστάδων στις οποίες οργανώνονται τα δεδομένα καθώς και των ορίων τους.

Οι Ultsch και Siemon [13], καθώς επίσης και ο Kraaijveld et al.[14] ανέπτυξαν έναν γραφικό τρόπο αναπαράστασης, καλούμενο U-matrix, για να επεξηγήσουν την ομαδοποίηση των διανυσμάτων αναφοράς στο SOM. Πρότειναν μια μέθοδο σύμφωνα με την οποία οι μέσες αποστάσεις μεταξύ των γειτονικών διανυσμάτων αναφοράς αντιπροσωπεύονται από αποχρώσεις του γκρι (εναλλακτικά μπορεί να χρησιμοποιηθεί μια κλίμακα χρωμάτων). Εάν η μέση απόσταση των γειτονικών mi είναι μικρή, χρησιμοποιείται μια ανοιχτή γκρι απόχρωση. Αντίστροφα, το σκούρο γκρι αντιπροσωπεύει τις μεγάλες αποστάσεις. Ο «χάρτης συστάδων» που διαμορφώνεται επί του SOM έπειτα, οπτικοποιεί με σαφήνεια την πραγματοποιούμενη ταξινόμηση . Ένα τέτοιο παράδειγμα ενός U-matrix απεικονίζεται στην εικόνα 2.5.



Εικόνα 2.5

2.11 Ερμηνεία της απεικόνισης που παράγει το SOM

Ο αυτοοργανούμενος χάρτης χάρτης συνδυάζει τις μεθόδους της μη γραμμικής προβολής κα της ομαδοποίησης σε έναν διατεταγμένο γράφο κβαντισμού διανυσμάτων. Επομένως, η χαρτογράφηση που παράγει αναμένεται να είναι εξηγήσιμη από την πλευρά ορισμένων κλασσικών εννοιών της στατιστικής.

Local Principal Components

Η κύρια διαφορά μεταξύ του SOM και της ανάλυσης κύριων συνιστωσών (PCA) είναι ότι η τελευταία περιγράφει τις συνολικές στατιστικές ιδιότητες της κατανομής των δεδομένων: ο πρώτος «κύριος άξονας» καθορίζει την κατεύθυνση στην οποία η μεταβλητότητα της κατανομής είναι η μεγαλύτερη, ο δεύτερος «κύριος άξονας» την κατεύθυνση όλων των κατευθύνσεων που είναι ορθογωνικές στον πρώτο κύριο άξονα στον οποίο το σφάλμα μεταβλητότητας είναι το μεγαλύτερο, και ούτω καθεξής.

Το SOM, εντούτοις, μπορεί να χαρακτηριστεί ως μια δισδιάστατη, πεπερασμένων στοιχείων «ελαστική επιφάνεια» ή ως ένα δίκτυο που περιγράφει ("εγκαθίσταται") την κατανομή των δειγμάτων εισόδου. Αυτό το δίκτυο έχει τοπικά, σε κάθε κόμβο του, δύο κύριες κατευθύνσεις που βρίσκονται θεωρώντας τις διαφορές μεταξύ των γειτονικών διανυσμάτων. Για παράδειγμα, το υποσύνολο των διανυσμάτων αναφοράς {mj} που αναπαριστά την πιο στενή γειτονιά του διανύσματος αναφοράς mi συμπεριλαμβανομένου του τελευταίου μπορεί να θεωρηθεί ως ένα ομαλοποιημένο (smooth) σύνολο δειγμάτων που προσπαθεί να αντιπροσωπεύσει μια τοπική δισδιάστατη υπερεπιφάνεια. Οι κύριοι άξονες αυτής της υπερεπιφάνειας και οι δύο μεγαλύτερες κύριες συνιστώσες αυτού του υποσυνόλου μπορούν να υπολογιστούν με τον κανονικό τρόπο από το υποσύνολο των γειτονικών διανυσμάτων αναφοράς.

Επειδή τα προκύπτοντα διανύσματα αναφοράς από τη διαδικασία κανονικοποίησης που εκτελεί το SOM μπορεί να μην αντιπροσωπεύουν με ακρίβεια τις πραγματικές στατιστικές ιδιότητες των δειγμάτων εισόδου, μπορούμε μόνο ποιοτικά να θεωρήσουμε τις δύο κύριες συνιστώσες ως "τοπικές κύριες συνιστώσες". Εάν θέλουμε περισσότερες από δύο, πρέπει να χρησιμοποιηθεί ένα υψηλότερης διάστασης SOM. Εν τούτοις, όπως οι «τοπικές κύριες συνιστώσες» μπορούν να υπολογιστούν εύκολα και με έναν σχεδόν μοναδικό τρόπο (που εξαρτάται μόνο από τη διαδικασία του SOM και τον καθορισμό της γειτονιάς), επιτρέπουν μιας επεξηγηματική όψη των σύνθετων και «θορυβωδών» κατανομών των δεδομένων.

Περαιτέρω μπορεί να εφαρμοστεί η κλασική ανάλυση παραγόντων σε κάθε σύνολο γειτονιάς των διανυσμάτων αναφοράς, προκειμένου να βρεθούν τα παραγοντικά φορτία των μεταβλητών εισόδου σε κάθε τοπική περιοχή του SOM. Βέβαια, τα τοπικά παραγοντικά φορτία μπορούν να είναι εντελώς διαφορετικά στις διαφορετικές περιοχές των δεδομένων εισόδου.

Σε αυτό το σημείο θα ήταν καλό να τονιστεί πως θα ήταν παράλογο να πραγματοποιηθεί μια τοπική PCA, ή να υπολογιστούν οι τοπικούς παράγοντες βάσει του υποσυνόλου των δειγμάτων εισόδου που απεικονίζονται στα γειτονικά σύνολα Voronoi. Επομένως είναι λογικότερο να καθοριστούν οι «τοπικοί παράγοντες» βάσει των διανυσμάτων αναφοράς. Με άλλα λόγια, όπως στην αποκαλούμενη διαφορική γεωμετρία, οι τοπικές κατευθύνσεις πρέπει να τηρήσουν ορισμένες «συνθήκες συμβατότητας» που το SOM λαμβάνει υπόψη αυτόματα.

Συμβολή μιας μεταβλητής στις δομές των συστάδων

Αυτή η ανάλυση σχετίζεται με την τοπική PCA και την τοπική ανάλυση παραγόντων αλλά είναι υπολογιστικά απλούστερη και περιγράφει αμεσότερα τη διακριτική δύναμη μιας μεταβλητής εισόδου στην απεικόνιση.

Αν εξετάσουμε της διανυσματικές διαφορές μεταξύ των γειτονικών διανυσμάτων αναφοράς, μπορούμε να διαμορφώσουμε ομοίως τις αντίστοιχες διαφορές κάθε διανυσματικής συνιστώσας χωριστά. Εάν η συσχέτιση μεταξύ των διανυσματικών διαφορών και οι διαφορές κάποιας συνιστώσας σε μια τοπική περιοχή του SOM είναι μεγάλες, αυτή η συνιστώσα (μεταβλητή) έχει μια σημαντική συμβολή στη δομή των συστάδων και μια μεγάλη επεξηγηματική δύναμη σε εκείνη την περιοχή των τιμών.

Η διακριτική δύναμη μιας μεταβλητής φανερώνεται εντονότερα στα σύνορα των συστάδων, όπου ψάχνει κανείς για τις μεταβλητές που προκαλούν τη μεγαλύτερη διαφορά μεταξύ των γειτονικών συστάδων.

Κεφάλαιο 3
Hidden Markov Models-HMM

3.1 Εισαγωγή

Οι διαδικασίες στον πραγματικό κόσμο παράγουν γενικά παρατηρήσιμα αποτελέσματα που μπορούν να χαρακτηριστούν ως σήματα. Τα σήματα μπορούν να είναι διακριτά στη φύση (πχ., χαρακτήρες ενός πεπερασμένου αλφάβητου, κβαντισμένα διανύσματα, κλπ.), ή συνεχή (πχ., δείγματα ομιλίας, μετρήσεις θερμοκρασίας, μουσική, κλπ.). Η πηγή των σημάτων μπορεί να είναι στάσιμη (δηλ., οι στατιστικές ιδιότητές της να μην ποικίλουν με το χρόνο), ή μη στάσιμη (δηλ., οι ιδιότητες σημάτων να ποικίλουν με την πάροδο του χρόνου). Τα σήματα μπορούν να είναι καθαρά (δηλ., προερχόμενα αυστηρά από μια πηγή), ή μπορούν να αλλοιώνονται από άλλες πηγές σημάτων (πχ., θόρυβος) ή από τις παραμορφώσεις μετάδοσης, την αντήχηση, κλπ.

Ένα πρόβλημα θεμελιώδους ενδιαφέροντος είναι ο χαρακτηρισμός τέτοιων σημάτων από την άποψη των μοντέλων σημάτων. Υπάρχουν αρκετοί λόγοι για τους οποίους μπορεί κανείς να ενδιαφέρεται για την εφαρμογή ενός τέτοιου μοντέλου. Καταρχήν, ένα μοντέλο σήματος μπορεί να παρέχει τη βάση για μια θεωρητική περιγραφή ενός συστήματος επεξεργασίας σήματος που μπορεί να χρησιμοποιείται για να επεξεργάζεται το σήμα ώστε να παρέχεται μια επιθυμητή έξοδος. Παραδείγματος χάριν, εάν ενδιαφερόμαστε για την ενίσχυση ενός σήματος λόγου που αλλοιώνεται από την παραμόρφωση θορύβου και την μετάδοση, μπορούμε να χρησιμοποιήσουμε μοντέλο σημάτων για να σχεδιάσουμε ένα σύστημα που θα αφαιρέσει με τον βέλτιστο τρόπο το θόρυβο και θα αναιρέσει την παραμόρφωση μετάδοσης. Ένας δεύτερος λόγος για τον οποίο τα μοντέλα σημάτων είναι σημαντικά είναι ότι είναι σε θέση να μας δώσουν ένα μεγάλο όγκο πληροφορίας για την πηγή του σήματος (δηλ., την πραγματική διαδικασία που παρήγαγε το σήμα) χωρίς να χρειάζεται να είναι διαθέσιμη η πηγή. Αυτή η ιδιότητα είναι ιδιαίτερα σημαντική όταν το κόστος λήψης των σημάτων από την πραγματική πηγή είναι υψηλό. Σε αυτήν την περίπτωση, με ένα καλό μοντέλο σημάτων, μπορούμε να προσομοιώσουμε την λειτουργία της πηγής και να μάθουμε όσο το δυνατόν περισσότερα μέσω των προσομοιώσεων. Τέλος, ο σημαντικότερος λόγος για τον οποίο τα μοντέλα σημάτων είναι σημαντικά είναι ότι συχνά λειτουργούν πάρα πολύ καλά στην πράξη και επιτρέπουν την πραγματοποίηση σημαντικών συστημάτων - πχ. συστήματα πρόβλεψης, συστήματα αναγνώρισης, συστήματα προσδιορισμού, κλπ., κατά τρόπο πολύ αποδοτικό.

Υπάρχουν διάφορες πιθανές επιλογές για το τί τύπος μοντέλου σήματος μπορεί να χρησιμοποιηθεί για το χαρακτηρισμό των ιδιοτήτων ενός δεδομένου σήματος. Χονδρικά, κάποιος μπορεί να ταξινομήσει τους τύπους μοντέλων σημάτων στις κλάσεις των αιτιοκρατικών μοντέλων και των στατιστικών μοντέλων. Τα αιτιοκρατικά μοντέλα εκμεταλλεύονται γενικά μερικές γνωστές συγκεκριμένες ιδιότητες του σήματος, πχ., ότι το σήμα είναι ένα κύμα ημιτόνου, ή ένα άθροισμα εκθετών, κλπ. Σε αυτές τις περιπτώσεις, η προδιαγραφή του μοντέλου σήματος είναι γενικά απλή και άμεση και το μόνο που απαιτείται είναι ο προσδιορισμός (υπολογισμός) των τιμών των παραμέτρων του μοντέλου(πχ. πλάτος, συχνότητα, φάση ενός κύματος ημιτόνου, πλάτη και ποσοστά εκθετών, κλπ.). Η δεύτερη μεγάλη κατηγορία μοντέλων σημάτων είναι το σύνολο των στατιστικών μοντέλων στα οποία κάποιος προσπαθεί να χαρακτηρίσει μόνο τις στατιστικές ιδιότητες του σήματος. Παραδείγματα τέτοιων στατιστικών μοντέλων περιλαμβάνουν τις γκαουσσιανές διαδικασίες, τις διαδικασίες Poisson, τις διαδικασίες Markov και τις hidden Markov διαδικασίες, μεταξύ άλλων. Η υπόθεση που γίνεται για το στατιστικό μοντέλο είναι ότι το σήμα μπορεί να χαρακτηριστεί ως μια παραμετρική τυχαία διαδικασία, και ότι οι παράμετροι της στοχαστικής διαδικασίας μπορούν να προσδιοριστούν (υπολογισθούν) κατά τρόπο ακριβή, καθορισμένο με σαφήνεια.

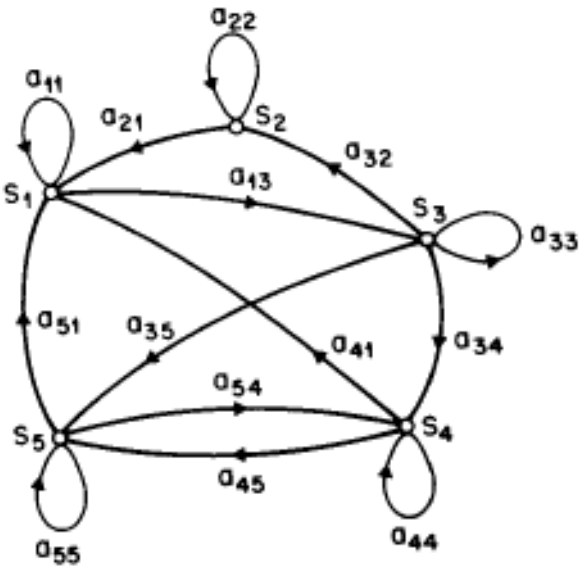
Στις επόμενες παραγράφους παρουσιάζεται αρχικά η θεωρία που αφορά τις αλυσίδες

Markov και στη συνέχεια τα Hidden Markov Models (HMM). Στη συνέχεια αναφέρονται τα τρία θεμελιώδη προβλήματα για το HMM, δηλαδή:

- αποτίμηση της πιθανότητας μιας ακολουθίας παρατηρήσεων δεδομένου ενός συγκεκριμένου HMM
- ο προσδιορισμός της καλύτερης ακολουθίας των καταστάσεων του μοντέλου
- η ρύθμιση των παραμέτρων του μοντέλου ώστε να περιγράψει καλύτερα το παρατηρηθέν σήμα.

Η θεωρία των HMM και των εφαρμογών της στην αναγνώριση λόγου παρουσιάζει μια ευρεία βιβλιογραφία. Η βασική θεωρία δημοσιεύθηκε σε μία σειρά κλασικών εργασιών από τον Baum και τους συναδέλφους του [2] - [6] προς το τέλος της δεκαετίας του '60 και στις αρχές της δεκαετίας του '70 και χρησιμοποιήθηκε σε εφαρμογές στην επεξεργασία του λόγου από τον Baker [7] στο CMU, και από τον Jelinek και τους συναδέλφους του στην IBM [7] - [13]. Εντούτοις, η διαδεδομένη κατανόηση και η εφαρμογή της θεωρίας των HMMs στην επεξεργασία του λόγου συντελέστηκε τη δεκαετία του '80 [1], [15] - [20].

3.2 Διακριτές διαδικασίες Markov



Εικόνα 3.1 Μαρκοβιανή αλυσίδα 5 καταστάσεων

Θεωρούμε ένα σύστημα που η κατάστασή του οποιαδήποτε χρονική στιγμή ανήκει σε ένα σύνολο N διακριτών καταστάσεων S1, S2…, SN. Σε διακριτές χρονικές στιγμές, το σύστημα υφίσταται μια αλλαγή κατάστασης (ενδεχομένως πίσω στην ίδια κατάσταση) σύμφωνα με ένα σύνολο πιθανοτήτων που αντιστοιχούν στην κατάσταση. Συμβολίζουμε τις χρονικές στιγμές που αντιστοιχούν σε αλλαγή κατάστασης με t=1, 2, ... και την πραγματική κατάσταση τη στιγμή t ως qt. Μια πλήρης πιθανοτική περιγραφή του ανωτέρω συστήματος, εν γένει, θα απαιτούσε τον προσδιορισμό της τρέχουσας κατάστασης (τη στιγμή t), καθώς επίσης και όλων των προηγούμενων καταστάσεων. Για την ειδική περίπτωση μιας διακριτής, πρώτης τάξης, αλυσίδας Markov, αυτή η πιθανοτική περιγραφή περιορίζεται αποκλειστικά στην παρούσα και την προηγούμενη κατάσταση, δηλ.,

$$P(q_t = s_j \mid q_{t-1} = s_i, \ldots, q_1 = s_k) = P(q_t = s_j \mid q_{t-1} = s_i) \quad (3.1)$$

Επιπλέον λαμβάνουμε υπόψιν μόνο εκείνες τις διαδικασίες στις οποίες η δεξιά πλευρά της (3.1) είναι ανεξάρτητη από το χρόνο, καταλήγοντας έτσι στο σύνολο των πιθανοτήτων μετάβασης κατάστασης aij, της μορφής

$$a_{ij} = P(q_t = s_j \mid q_{t-1} = s_i), \quad t>1, \quad s_i, s_j \in S \quad (3.2)$$

με τους συντελεστές μετάβασης κατάστασης να έχουν τις ιδιότητες aij ≥ 0 και

$$\sum_{j=1}^{N} a_{ij} = 1$$

δεδομένου ότι υπακούν τους τυπικούς στοχαστικούς περιορισμούς.

Η ανωτέρω στοχαστική διαδικασία θα μπορούσε να ονομαστεί ένα παρατηρήσιμο μοντέλο Markov δεδομένου ότι η έξοδος της διαδικασίας είναι ένα σύνολο καταστάσεων κάθε χρονική στιγμή, με κάθε κατάσταση να αντιστοιχεί σε ένα φυσικό (παρατηρήσιμο) γεγονός. Για παράδειγμα έστω ένα απλό μαρκοβιανό μοντέλο του καιρού. Υποθέτουμε ότι μια φορά την ημέρα (πχ, το μεσημέρι), ο καιρός παρατηρείται να βρίσκεται σε μια από τις εξής καταστάσεις:
κατάσταση 1: βροχή ή (χιόνι)
κατάσταση 2: συννεφιά
κατάσταση 3: ηλιοφάνεια.

Υποθέτοντας ότι ο καιρός την ημέρα t χαρακτηρίζεται από μια μόνο από τις παραπάνω τρεις καταστάσεις, και ότι η μήτρα Α των πιθανοτήτων μετάβασης κατάστασης είναι

$$A = \begin{matrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{matrix}$$

Δεδομένου ότι ο καιρός την ημέρα 1 (t = 1) είναι ηλιόλουστος (κατάσταση 3), μπορούμε να υποβάλουμε την ερώτηση: Ποια είναι η πιθανότητα (σύμφωνα με το μοντέλο) ο καιρός για τις επόμενες 7 ημέρες να είναι «ήλιος-ήλιος-βροχή-βροχή-ήλιος-συννεφιά-ήλιος....»; Τυπικότερα, ορίζουμε την ακολουθία παρατηρήσεων Ο ως *O = {S3, S3, S3, S1, S1, S3, S2, S3}* με αντίστοιχους χρόνους t = 1, 2,…, 8, και επιθυμούμε να καθορίσουμε την πιθανότητα της Ο να συμβεί, δοθέντος του μοντέλου. Αυτή η πιθανότητα μπορεί να εκφραστεί (και να υπολογιστεί) ως εξής:

$$P(O \mid Model) = \Pi[S3, S3, S3, S1, S1, S3, S2, S3 \mid Model] =$$
$$= P[S3]\cdot P[S3 \mid S3]\cdot P[S3 \mid S3]\cdot P[S1 \mid S3]\cdot P[S1 \mid S1]\cdot P[S3 \mid S1]\cdot P[S2 \mid S3]\cdot P[S3 \mid S2] =$$
$$= 1(0.8)(0.8)(0.1)(0.4)(0.3)(0.1)(0.2) = 1.536 \times 10^{-4}$$

Μια άλλη ενδιαφέρουσα ερώτηση μπορούμε να απαντήσουμε χρησιμοποιώντας το μοντέλο είναι: Δεδομένου ότι το μοντέλο είναι σε μία γνωστή κατάσταση, ποια είναι η πιθανότητα να παραμείνει σε αυτή την κατάσταση για ακριβώς d ημέρες;
Αυτή η πιθανότητα μπορεί να αποτιμηθεί ως η πιθανότητα της ακολουθίας παρατηρήσεων

$$O = \left\{ S_{i_1}, S_{i_2}, S_{i_1} \ldots S_{i_d}, S_{i_{d+1}} \neq S_i \right\}$$

δοθέντος του μοντέλου, που είναι

P(O|Model, ql = Si) = $(aii)^{d-1}$ (l - aii) = pi(d)

Η ποσότητα pi(d) είναι η διακριτή συνάρτηση πυκνότητας πιθανότητας της παραμονής d στην κατάσταση i. Αυτή η εκθετική πυκνότητα παραμονής είναι χαρακτηριστική της παραμονής σε μια κατάσταση σε μια αλυσίδα Markov. Με βάση την pi(d), μπορούμε εύκολα να υπολογίσουμε τον αναμενόμενο αριθμό παρατηρήσεων (παραμονή) σε μία κατάσταση, υποθέτοντας ότι ξεκινάμε από αυτή την κατάσταση, ως

$$\overline{d_i} = \sum_{d=1}^{\infty} d\rho_i(d) = \sum_{d=1}^{\infty} d(a_{ii})^{d-1}(1-a_{ii}) = \frac{1}{1-a_{ii}} \qquad (3.3)$$

Κατά συνέπεια ο αναμενόμενος αριθμός διαδοχικών ημερών του ηλιοφάνειας, σύμφωνα με το μοντέλο, είναι 1/0.2 = 5 για νεφελώδη καιρό είναι 2.5 και για τη βροχερό είναι 1.67.

3.3 Hidden Markov Models

Μέχρι τώρα έχουμε εξετάσει τα μαρκοβιανά μοντέλα στα οποία κάθε κατάσταση αντιστοιχούσε σε ένα παρατηρήσιμο (φυσικό) γεγονός. Αυτό το μοντέλο είναι πάρα πολύ περιοριστικό για να εφαρμοστεί σε πολλά προβλήματα που μας ενδιαφέρουν. Σε αυτή την ενότητα επεκτείνουμε την έννοια των μαρκοβιανών μοντέλων ώστε να περιλάβουν την περίπτωση όπου η παρατήρηση είναι μια πιθανοτική συνάρτηση της κατάστασης δηλαδή, το προκύπτον μοντέλο (που είναι ένα hidden Markov model) είναι μια διπλά ενσωματωμένη στοχαστική διαδικασία με μια ελλοχεύουσα στοχαστική διαδικασία που δεν είναι παρατηρήσιμη(είναι κρυμμένη), αλλά μπορεί μόνο να παρατηρηθεί μέσω ενός άλλου συνόλου στοχαστικών διαδικασιών που παράγουν την ακολουθία παρατηρήσεων.

3.3.1 Δύο απλά παραδείγματα

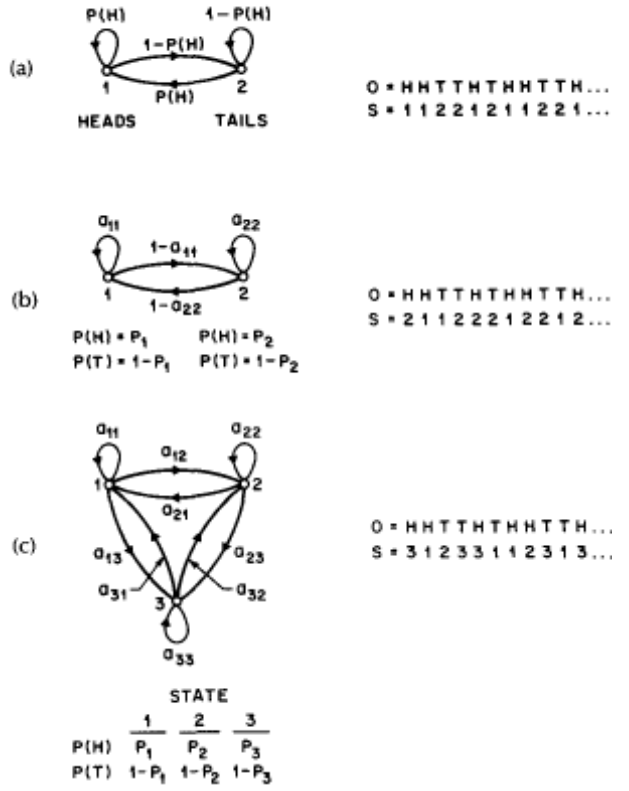
Μοντέλο ρίψης νομισμάτων: Υποθέστε το ακόλουθο σενάριο. Βρισκόμαστε σε ένα δωμάτιο με ένα εμπόδιο (πχ, μια κουρτίνα) μέσω του οποίου δεν μπορούμε να δούμε τι συμβαίνει. Στην άλλη πλευρά του εμποδίου είναι ένα άλλο άτομο που εκτελεί ένα πείραμα ρίψης νομισμάτων . Το άλλο άτομο δεν μας ενημερώνει για το τί ακριβώς κάνει, θα μας πει μόνο το αποτέλεσμα κάθε ρίψης νομισμάτων. Κατά συνέπεια, μια ακολουθία κρυμμένων πειραμάτων ρίψης νομισμάτων εκτελείται, με την ακολουθία παρατηρήσεων να αποτελείται από μια σειρά από κεφαλές και γράμματα:
πχ μια τυπική ακολουθία παρατηρήσεων θα μπορούσε να είναι
O= O1 O2 O3....OT = ΚΚΓΓΓΚΓΓΚ...Κ

Λαμβάνοντας υπόψη το ανωτέρω σενάριο, το πρόβλημα που μας ενδιαφέρει είναι πώς κατασκευάζουμε ένα HMM για να εξηγήσουμε (μοντέλο) την παρατηρηθείσα ακολουθία κεφαλών και γραμμάτων. Το πρώτο πρόβλημα που κανείς συναντά είναι να αποφασίσει σε τι αντιστοιχούν οι καταστάσεις του μοντέλου και έπειτα πόσες καταστάσεις πρέπει να υπάρχουν. Μια πιθανή επιλογή είναι να υποτεθεί πως μόνο ένα "πολωμένο" νόμισμα έχει ριφθεί. Σε αυτήν την περίπτωση, θα μπορούσαμε να μοντελοποιήσουμε την κατάσταση με ένα μοντέλο δύο καταστάσεων όπου κάθε κατάσταση αντιστοιχεί σε μια πλευρά του νομίσματος (δηλαδή, κεφαλή ή γράμματα). Αυτό το μοντέλο απεικονίζεται στην εικόνα 3.2(α). Σε αυτήν την περίπτωση, το μοντέλο Markov είναι παρατηρήσιμο, και το μόνο

ζήτημα για την πλήρη προδιαγραφή του μοντέλου θα ήταν να αποφασιστεί σχετικά με την καλύτερη τιμή για την πόλωση (δηλ., η πιθανότητα, για παράδειγμα, των κεφαλών). Ένα ισοδύναμο HMM θα ήταν ένα εκφυλισμένο μοντέλο μίας κατάστασης, όπου η κατάσταση αντιστοιχεί στο ενιαίο πολωμένο νόμισμα, και η άγνωστη παράμετρος είναι η πόλωση του νομίσματος.

Μια δεύτερη μορφή HMM για την να εξηγήσουμε την παρατηρηθείσα ακολουθία ρίψεων νομισμάτων δίνεται στην εικόνα 3.2 (β). Σε αυτήν την περίπτωση υπάρχουν 2 καταστάσεις στο μοντέλο και κάθε κατάσταση αντιστοιχεί σε έναν διαφορετικό, πολωμένο, νόμισμα που πετιέται. Κάθε κατάσταση χαρακτηρίζεται από μια κατανομή πιθανότητας των κεφαλών και των γραμμάτων, και οι μεταβάσεις μεταξύ των καταστάσεων χαρακτηρίζονται από μια μήτρα μετάβασης καταστάσεων. Ο φυσικός μηχανισμός που αποφασίζει το πώς επιλέγονται οι μεταβάσεις καταστάσεων θα μπορούσε ο ίδιος να είναι ένα σύνολο ανεξάρτητων ρίψεων νομισμάτων ή κάποιο άλλο πιθανοτικό γεγονός.

Μια τρίτη μορφή HMM για την εξήγηση της παρατηρηθείσας ακολουθίας ρίψεων νομισμάτων δίνεται στην εικόνα 3.2(γ). Αυτό το μοντέλο αντιστοιχεί στη χρησιμοποίηση 3 πολωμένων νομισμάτων, και την επιλογή μεταξύ των τριών, με βάση κάποιο πιθανοτικό γεγονός

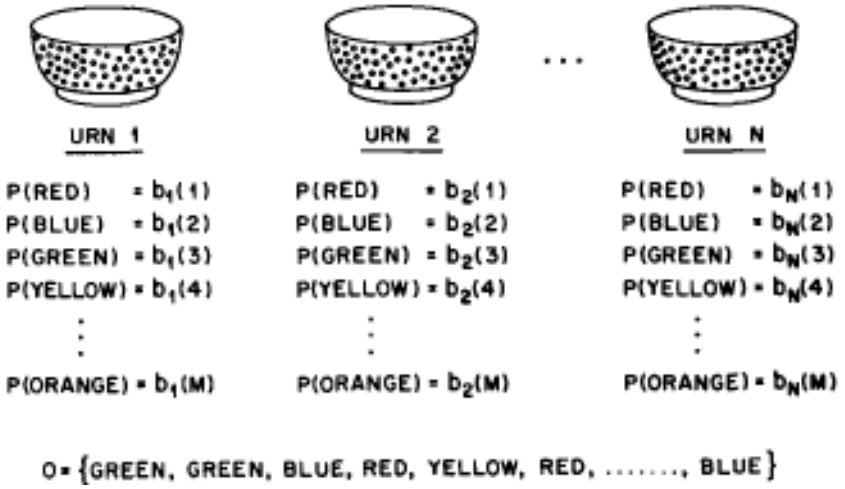


Εικόνα 3.2 (α) μοντέλο 1 νομίσματος (β) μοντέλο 2 νομισμάτων (γ) για 3 νομίσματα

Λαμβάνοντας υπόψη την επιλογή μεταξύ των τριών μοντέλων που παρουσιάζονται στην εικόνα 3.2 για την εξήγηση της παρατηρηθείσας ακολουθίας κεφαλών και γραμμάτων, μια αναμενόμενη ερώτηση θα ήταν ποιο μοντέλο ταιριάζει καλύτερα με τις υπάρχουσες παρατηρήσεις. Είναι φανερό πως το απλό μοντέλο 1 νομίσματος έχει μόνο μία άγνωστη παράμετρο το μοντέλο 2 νομισμάτων έχει άγνωστες 4 παραμέτρους και το μοντέλο 3 νομισμάτων έχει 9 άγνωστες παραμέτρους. Κατά συνέπεια, έχοντας τους περισσότερους

βαθμούς ελευθερίας, το μεγαλύτερο HMM θα φαινόταν να είναι εγγενώς ικανότερο να μοντελοποιήσει μια σειρά από πειράματα ρίψης νομισμάτων από ότι ισοδύναμα μικρότερα μοντέλα. Αν και αυτό ισχύει θεωρητικά, πρακτικοί λόγοι επιβάλλουν μερικούς ισχυρούς περιορισμούς στο μέγεθος των μοντέλων που μπορούμε να εξετάσουμε. Επιπλέον, μπορεί να βρισκόμαστε στην περίπτωση όπου ακριβώς μόνο ένα νόμισμα ρίπτεται. Σε αυτή την περίπτωση, η χρησιμοποίηση του μοντέλου 3 νομισμάτων θα ήταν ακατάλληλη, δεδομένου ότι το πραγματικό γεγονός δεν θα αντιστοιχούσε στο μοντέλο που χρησιμοποιείται δηλαδή, θα χρησιμοποιούσαμε ένα υποκαθορισμένο σύστημα.

Μοντέλο δοχείων και σφαιρών: Ένα πιο σύνθετο σύστημα αποτελεί το σύστημα δοχείων και σφαιρών της εικόνας 3.3. Υποθέτουμε ότι υπάρχουν N μεγάλα γυάλινα δοχεία. Σε κάθε δοχείο βρίσκεται ένας μεγάλος αριθμός χρωματισμένων σφαιρών. Υποθέτουμε ότι υπάρχουν M ευδιάκριτα χρώματα σφαιρών. Η διαδικασία λήψης παρατηρήσεων είναι η ακόλουθη. Ένα άτομο βρίσκεται στο δωμάτιο, και σύμφωνα με κάποια τυχαία διαδικασία, επιλέγει ένα αρχικό δοχείο. Από αυτό το δοχείο μια σφαίρα επιλέγεται τυχαία, και το χρώμα της καταγράφεται ως παρατήρηση. Η σφαίρα αντικαθίσταται έπειτα στο δοχείο από το οποίο επιλέχτηκε. Ένα νέο δοχείο επιλέγεται έπειτα σύμφωνα με την τυχαία διαδικασία εκλογής που συνδέεται με το τρέχον δοχείο, και η διαδικασία επιλογής σφαίρας επαναλαμβάνεται. Αφού ολοκληρωθεί η διαδικασία αυτή λαμβάνουμε μία πεπερασμένη ακολουθία παρατηρήσεων χρωμάτων, την οποία θα επιθυμούσαμε να μοντελοποιήσουμε ως αισθητή παρατηρήσιμη έξοδο ενός HMM.



Εικόνα 3.3 μοντέλο δοχείων-σφαιρών

Είναι προφανές ότι το απλούστερο HMM που αντιστοιχεί στη ανωτέρω διαδικασία δοχείων και σφαιρών είναι ένα στο οποίο κάθε κατάσταση αντιστοιχεί σε ένα συγκεκριμένο δοχείο, και για το οποίο ορίζεται μια πιθανότητα χρώματος (σφαίρας) για κάθε κατάσταση. Η επιλογή των δοχείων είναι υπαγορευμένη από τη μήτρα μετάβασης καταστάσεων του HMM.

3.3.2 Παράμετροι ενός HMM

Τα ανωτέρω παραδείγματα μας δίνουν μια αρκετά καλή ιδέα για αυτό που ένα HMM είναι και πώς μπορεί να εφαρμοστεί σε μερικά απλά σενάρια. Σε αυτή την παράγραφο, ορίζουμε με τυπικό τρόπο τα στοιχεία ενός HMM και εξηγούμε πώς το μοντέλο παράγει τις

ακολουθίες παρατηρήσεων.
Ένα HMM χαρακτηρίζεται από τα εξής:

1) N, ο αριθμός καταστάσεων στο μοντέλο. Αν και οι καταστάσεις είναι κρυμμένες, για πολλές πρακτικές εφαρμογές υπάρχει συχνά κάποια φυσική σημασία που συνδέεται με τις καταστάσεις ή με τα σύνολα καταστάσεων του μοντέλου. Ως εκ τούτου, στα πειράματα ρίψης νομισμάτων, κάθε κατάσταση αντιστοιχούσε σε ένα ευδιάκριτο πολωμένο νόμισμα. Στο μοντέλο δοχείων και σφαιρών, οι καταστάσεις αντιστοιχούσαν στα δοχεία. Γενικά, οι καταστάσεις διασυνδέονται κατά τέτοιο τρόπο ώστε οποιοδήποτε κατάσταση μπορεί να επιτευχθεί από οποιαδήποτε άλλη κατάσταση (πχ, ένα εργοδικό μοντέλο) εντούτοις, και άλλες πιθανές διασυνδέσεις των καταστάσεων είναι συχνά ενδιαφέροντες. Συμβολίζουμε τις μεμονωμένες καταστάσεις ως S = {S1, S2,…, SN}, και την κατάσταση τη χρονική t ως qt.

2) M, ο αριθμός διακριτών συμβόλων παρατηρήσεων ανά κατάσταση, δηλαδή, το μέγεθος του αλφάβητου. Τα σύμβολα των παρατηρήσεων αντιστοιχούν στην έξοδο του συστήματος που μοντελοποιείται. Για τα πειράματα ρίψεων νομισμάτων τα σύμβολα παρατηρήσεων ήταν απλά κεφαλές ή γράμματα, για το μοντέλο σφαιρών και δοχείων ήταν τα χρώματα των σφαιρών που επιλέχτηκαν από τα δοχεία. Συμβολίζουμε τα μεμονωμένα σύμβολα ως V = {vl, v2, ...,vM}.

3) Η κατανομή πιθανότητας μετάβασης κατάστασης A = {aij} όπου

$$\mathbf{aij = P[q_{t+1} = Sj \mid qt = Si],\ 1 \le i, j \le N} \qquad \mathbf{(3.4)}$$

Για την ειδική περίπτωση όπου οποιαδήποτε κατάσταση μπορεί να φθάσει σε οποιαδήποτε άλλη κατάσταση σε ένα βήμα, ισχύει aij > 0 για όλα τα i, j. Για άλλους τύπους HMMs, θα ίσχυε aij = 0 για ένα ή περισσότερα ζευγάρια (i, j).

4) Η κατανομή πιθανότητας των συμβόλων παρατηρήσεων στην κατάσταση j,

B={bj(k)}, όπου $\mathbf{bj(k) = P\ [\ vk\ at\ t \mid qt = Sj],\ 1 \le j \le N,\ 1 \le k \le M}$ **(3.5)**

5) Η κατανομή αρχικής κατάστασης π= {πi} όπου $\mathbf{\pi i = P\ [ql = Si],\ 1 \le i \le N}$ **(3.6)**

Δοθέντων κατάλληλων τιμών για τα N, M, A, B, και π, το HMM μπορεί να χρησιμοποιηθεί ως γεννήτρια μιας ακολουθίας παρατηρήσεων O= O1 O2.... OT (όπου κάθε παρατήρηση Ot, είναι ένα από τα σύμβολα του V, και T είναι ο αριθμός παρατηρήσεων στην ακολουθία) ως εξής:

1) Επέλεξε μια αρχική κατάσταση q1 = Si σύμφωνα με τη κατανομή αρχικής κατάστασης π.
2) Θέσε t = 1
3) Επέλεξε Ot = vk σύμφωνα με τη κατανομή πιθανότητας συμβόλων στην κατάσταση Si, δηλαδή, bi(k).
4) Μετάβαση σε μια νέα κατάσταση q_{t+1} = Sj σύμφωνα με τη κατανομή πιθανότητας μετάβασης κατάστασης για την κατάσταση Sj δηλαδή aij.
5) Θέσε t = t + 1 και επιστροφή στο βήμα 3) εάν t < T διαφορετικά ολοκληρώστε τη διαδικασία.

Η ανωτέρω διαδικασία μπορεί να χρησιμοποιηθεί ως γεννήτρια παρατηρήσεων αλλά και ως μοντέλο για το πώς μια δεδομένη παρατήρηση η ακολουθία παρήχθη από ένα

κατάλληλο HMM.

Γίνεται φανερό από τα παραπάνω ότι μια πλήρης προδιαγραφή ενός HMM απαιτεί την προδιαγραφή δύο παραμέτρων του μοντέλου (N και M), προδιαγραφή των συμβόλων παρατηρήσεων, και προδιαγραφή των τριών κατανομών πιθανοτήτων A, B, και π. Για ευκολία, χρησιμοποιούμε τη σύμβαση

$$\lambda=(A,B,\pi) \quad (3.7)$$

για να συμβολίσουμε το πλήρες σύνολο παραμέτρων του μοντέλου.

3.3.3 Τα τρία θεμελιώδη προβλήματα για HMMs

Λαμβάνοντας υπόψη τη μορφή του HMM της προηγούμενης παραγράφου, υπάρχουν τρία βασικά προβλήματα που μας ενδιαφέρει να επιλυθούν ώστε το μοντέλο να είναι χρήσιμο στις εφαρμογές του πραγματικού κόσμου. Αυτά τα προβλήματα είναι τα ακόλουθα:

Πρόβλημα 1: Λαμβάνοντας υπόψη την ακολουθία των παρατηρήσεων O= O1 O2...OT, και ένα μοντέλο λ = (A, B,π), πώς υπολογίζουμε αποτελεσματικά την P(O|λ), την πιθανότητα εμφάνισης της ακολουθίας O δοθέντος του μοντέλου λ;

Πρόβλημα 2: Λαμβάνοντας υπόψη την ακολουθία των παρατηρήσεων O= O1 O2...OT, και το μοντέλο λ, πώς επιλέγουμε μια αντίστοιχη ακολουθία καταστάσεων Q = q1 q2…qT που είναι βέλτιστη υπό κάποια έννοια (δηλαδή, 'εξηγεί' καλύτερα τις παρατηρήσεις);

Πρόβλημα 3: Πώς ρυθμίζουμε τις παραμέτρους του μοντέλου λ = (A, B, π) για να μεγιστοποιήσουμε την P(O|λ);

Το πρόβλημα 1 είναι το πρόβλημα της αποτίμησης (evaluation), δηλαδή λαμβάνοντας υπόψη ένα μοντέλο και μια ακολουθία από παρατηρήσεις, πώς υπολογίζουμε την πιθανότητα η παρατηρηθείσα ακολουθία να παρήχθη από το μοντέλο. Μπορούμε επίσης να δούμε το πρόβλημα όπως μια βαθμολόγηση του πόσο καλά ένα δεδομένο μοντέλο ταιριάζει με μια δεδομένη ακολουθία παρατηρήσεων. Η τελευταία παρατήρηση είναι εξαιρετικά χρήσιμη. Παραδείγματος χάριν, εάν εξετάζουμε την περίπτωση στην οποία προσπαθούμε να επιλέξουμε μεταξύ διάφορων ανταγωνιστικών μοντέλων, η λύση στο πρόβλημα 1 μας επιτρέπει να επιλέξουμε το μοντέλο που ταιριάζει καλύτερα στις παρατηρήσεις.

Το πρόβλημα 2 είναι αυτό στο οποίο προσπαθούμε να αποκαλύψουμε το κρυμμένο μέρος του μοντέλου, δηλαδή, για να βρούμε τη «σωστή» ακολουθία καταστάσεων. Πρέπει να γίνει σαφές ότι για όλες τις περιπτώσεις εκτός από αυτές των εκφυλισμένων μοντέλων, δεν υπάρχει καμία «σωστή» ακολουθία καταστάσεων. Ως εκ τούτου για τις πρακτικές περιπτώσεις, χρησιμοποιούμε συνήθως ένα κριτήριο βελτιστοποίησης για να λύσουμε αυτό το πρόβλημα όσο το δυνατόν καλύτερα. Δυστυχώς, υπάρχουν διάφορα κριτήρια βελτιστοποίησης που μπορούν να εφαρμοστούν και ως εκ τούτου η επιλογή του κριτηρίου εξαρτάται στενά της προοριζόμενης χρήσης για την παρατηρούμενη ακολουθία καταστάσεων. Χαρακτηριστικές χρήσεις μπορεί να είναι να θέλουμε να μάθουμε τη δομή του μοντέλου, να βρούμε τις βέλτιστες ακολουθίες καταστάσεων για αναγνώριση φωνής ή να πάρουμε τις μέσες στατιστικές τιμές των μεμονωμένων καταστάσεων κλπ.

Το πρόβλημα 3 είναι αυτό στο οποίο προσπαθούμε να βελτιστοποιήσουμε τις παραμέτρους του μοντέλου ώστε αυτό να περιγραφεί καλύτερα πώς μια δεδομένη ακολουθία παρατηρήσεων παράγεται. Η ακολουθία παρατηρήσεων που χρησιμοποιείται για να ρυθμίσει τις παραμέτρους του μοντέλου καλείται ακολουθία εκπαίδευσης δεδομένου ότι χρησιμοποιείται για «να εκπαιδεύσει» το HMM. Το πρόβλημα εκπαίδευσης είναι το κρίσιμο για τις περισσότερες εφαρμογές HMMs, δεδομένου ότι μας επιτρέπει να προσαρμόσουμε

βέλτιστα τις παραμέτρους του μοντέλου στα παρατηρηθέντα δεδομένα εκπαίδευσης δηλαδή, να δημιουργήσουμε τα καλύτερα μοντέλα για φαινόμενα του πραγματικού κόσμου.

3.4 Λύσεις των τριών θεμελιωδών προβλημάτων των HMMs

3.4.1 Πρόβλημα 1

Επιθυμούμε να υπολογίσουμε την πιθανότητα εμφάνισης της ακολουθίας παρατηρήσεων, O=O1 O2…. OT, δοθέντος του μοντέλου λ, δηλαδή, P(O|λ). Ο απλούστερος τρόπος υπολογισμού είναι μέσω της απαρίθμησης κάθε πιθανής κατάστασης ακολουθίας μήκους T (ο αριθμός παρατηρήσεων). Θεωρούμε μια τέτοια σταθερή ακολουθία καταστάσεων Q = q 1,q 2...qT όπου q1 είναι η αρχική κατάσταση. Η πιθανότητα εμφάνισης της ακολουθίας O για την ακολουθία Q είναι

$$\mathbf{P(O|Q, \lambda) =} \prod_{t=1}^{T} P(O_t|q_t, \lambda) \qquad \textbf{(3.8)}$$

όπου έχουμε υποθέσει στατιστική ανεξαρτησία των παρατηρήσεων και άρα

$$\mathbf{P(O|Q, \lambda) =} \; b_{q1}(O_1)b_{q2}(O_2)\ldots b_{qT}(O_T) \qquad \textbf{(3.9)}$$

Η πιθανότητα μιας τέτοιας ακολουθίας καταστάσεων Q μπορεί να γραφτεί ως

$$\mathbf{P(Q|\lambda) =} \; \pi_{q1}a_{q1q2}a_{q2q3}\ldots a_{qT-1qT} \qquad \textbf{(3.10)}$$

Η από κοινού πιθανότητα των O και Q, δηλαδή η πιθανότητα ότι τα O και Q συμβαίνουν ταυτόχρονα, είναι απλά το γινόμενο των ανωτέρω δύο όρων ήτοι,

$$\mathbf{P(O, Q|\lambda) = P(O|Q, \lambda)P(Q, \lambda)}$$

Η πιθανότητα O (δοθέντος του μοντέλου) λαμβάνεται αθροίζοντας αυτή την από κοινού πιθανότητα σε όλες τις πιθανές ακολουθίες καταστάσεων q. Τελικά,

$$\mathbf{P(O|\lambda) =} \sum_{all\,Q} \pi_{q1}b_{q1}(O_1)a_{q1q2}b_{q2}(O_2)\ldots a_{qT-1qT}b_{qT}(O_T) \qquad \textbf{(3.11)}$$

Η ερμηνεία του υπολογισμού στην ανωτέρω εξίσωση είναι η ακόλουθη. Αρχικά (τη χρονική στιγμή t = 1) το σύστημα βρίσκεται στην κατάσταση q1 με την πιθανότητα π_{q1} και παράγει το σύμβολο O1 (σε αυτή την κατάσταση) με πιθανότητα b_{q1}(O1). Στην επόμενη χρονική στιγμή t+1 (t = 2) μεταβαίνει στην κατάσταση q2, από την κατάσταση q1 με πιθανότητα a_{q1q2}, και παράγει το σύμβολο O2 με πιθανότητα b_{q2}(O2). Αυτή η διαδικασία συνεχίζεται κατά αυτόν τον τρόπο έως μεταβούμε στην κατάσταση qT από την κατάσταση qT-1 (τη χρονική στιγμή T) με πιθανότητα a_{qT-1qT} και παραχθεί το σύμβολο OT με πιθανότητα b_{qT} (OT).

Είναι εύκολο να δούμε πως ο υπολογισμός της P(O|λ) σύμφωνα με τον άμεσο ορισμό της εξίσωσης (3.11) απαιτεί υπολογισμούς της τάξης 2Tx N^T, δεδομένου ότι κάθε χρονική

στιγμή t = 1, 2,…, T, υπάρχουν N καταστάσεις που μπορούν να επιτευχθούν (δηλαδή, υπάρχουν N^T πιθανές ακολουθίες καταστάσεων), και για κάθε τέτοια ακολουθία καταστάσεων απαιτούνται περίπου 2T υπολογισμοί για κάθε όρο στην (3.11). Αυτός ο υπολογισμός είναι υπολογιστικά ανέφικτος, ακόμη και για μικρές τιμές των N και T πχ, για N = 5 (καταστάσεις), T = 100 (παρατηρήσεις), είναι της τάξης 2x100 x5^{100}= 10^{72} υπολογισμοί! Μια αποδοτικότερη διαδικασία είναι η forward-backward procedure.

H Forward-Backward procedure [3], [4]: Θεωρούμε την μεταβλητή αt(i) που ορίζεται ως **αt(i)= P(O1,O2… Ot, qt = Si|λ)** δηλαδή, η πιθανότητα εμφάνισης της μερικής ακολουθίας παρατηρήσεων, O1, O2...Ot, (μέχρι την χρονική στιγμή t) και της κατάστασης Si τη χρονική στιγμή t, δοθέντος του μοντέλου λ. Μπορούμε να προσδιορίσουμε το αt(i) αναδρομικά ως εξής:

1) Αρχικοποίηση:

$$\alpha_1(i) = \pi_i b_1(O_1), \qquad 1 \leqslant i \leqslant N \tag{3.12α}$$

2) Αναδρομή:

$$\alpha_{t+1}(j) = \left[\sum_{j=1}^{N} \alpha_t(i)\alpha_{ij}\right] b_j(O_{t+1}), \qquad 1 \leqslant j \leqslant N,\ 1 \leqslant t \leqslant T-1 \tag{3.12β}$$

3) Τερματισμός:

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{3.12γ}$$

Το πρώτο βήμα αρχικοποιεί τις "προς τα εμπρός" πιθανότητες της κατάστασης Si και της αρχικής παρατήρησης O1. Το βήμα της αναδρομής, αποτελεί τον πυρήνα του αλγορίθμου. Μέσω αυτού υπολογίζεται η πιθανότητα να βρεθούμε στην κατάσταση Sj από οιαδήποτε από τις N καταστάσεις, τη χρονική στιγμή t+1 και έχοντας παρατηρήσει την υποακολουθία O1...Ot. Ο υπολογισμός αυτός υλοποιείται για όλες τις καταστάσεις για δοσμένη χρονική στιγμή t και επαναληπτικά για όλες τις χρονικές στιγμές ως T. Τέλος, το βήμα 3 δίνει τον επιθυμητό υπολογισμό της P(O|λ) ως το άθροισμα των τελικών forward μεταβλητών $\alpha_T(i)$.

Γίνεται εύκολα αντιληπτό ότι οι απαιτούμενοι υπολογισμοί τώρα πλέον είναι της τάξης N^2T, αρκετές τάξεις μεγέθους μικρότεροι σε σχέση με τον άμεσο υπολογισμό της πιθανότητας.

Κατά παρόμοιο τρόπο μπορούμε να ορίσουμε μια "προς τα πίσω" (backward) μεταβλητή βt(i) ως

$$\beta_t(i) = P(O_{t+1}O_{t+2}\ldots O_T|q_t = S_i, \lambda)$$

δηλαδή, την πιθανότητα η μερική ακολουθία παρατηρήσεων από την χρονική στιγμή t+1 ως το τέλος, δοθείσης της κατάστασης Si τη χρονική στιγμή t και το μοντέλο λ. Ομοίως με πριν, μπορούμε να λύσουμε αναδρομικά ως εξής:

1) Αρχικοποίηση:

$$\beta_T(i) = 1 \qquad 1 \leqslant i \leqslant N \tag{3.13α}$$

2) Επαγωγή:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j) \qquad 1 \leqslant i \leqslant N \quad \text{t=T-1, T-2,..1} \tag{3.13β}$$

Το βήμα 1 ορίζει αυθαίρετα το $\beta_T(i)$ να είναι 1 για όλα τα i. Το δεύτερο βήμα, δείχνει

ότι προκειμένου να βρισκόμαστε στην κατάσταση Si τη χρονική στιγμή t, και για να παρατηρηθεί η ακολουθία παρατηρήσεων από τη χρονική στιγμή t+1 και έπειτα, πρέπει να ληφθούν υπόψιν όλες οι πιθανές καταστάσεις Sj τη χρονική στιγμή t+1, πραγματοποιώντας τη μετάβαση από το Si στο Sj (ο όρος aij), καθώς επίσης και η παρατήρηση O_{t+1} στην κατάσταση j (ο όρος β_{t+1} (j)), και έπειτα να λάβουμε υπόψιν την υπόλοιπη μερική ακολουθία παρατηρήσεων από την κατάσταση j. Οι παραπάνω υπολογισμοί (forward, backward) χρησιμοποιούνται εκτενώς στην επίλυση των άλλων δύο θεμελιωδών προβλημάτων των HMMs.

3.4.2 Πρόβλημα 2

Αντίθετα με το πρόβλημα 1 για το οποίο μπορεί να δοθεί μια ακριβής λύση, υπάρχουν διάφοροι πιθανοί τρόποι επίλυσης για το πρόβλημα 2, δηλαδή την εύρεση της «βέλτιστης» ακολουθίας καταστάσεων που συνδέεται με τη δεδομένη ακολουθία παρατηρήσεων. Η δυσκολία έγκειται στον ορισμό της βέλτιστης ακολουθίας καταστάσεων αφού υπάρχουν διάφορα πιθανά κριτήρια βελτιστότητας. Παραδείγματος χάριν, ένα τέτοιο πιθανό κριτήριο είναι να επιλεχθούν οι καταστάσεις q, οι οποίες είναι μεμονωμένα πιο πιθανές. Αυτό το κριτήριο μεγιστοποιεί τον αναμενόμενο αριθμό σωστών μεμονωμένων καταστάσεων. Για να εφαρμόσουμε αυτήν την λύση στο πρόβλημα 2, ορίζουμε τη μεταβλητή

$$\gamma_t(i) = P(q_t = S_i | O, \lambda) \tag{3.14}$$

ήτοι την πιθανότητα να βρίσκεται το σύστημα στην κατάσταση Si τη χρονική στιγμή t, λαμβάνοντας υπόψη την ακολουθία των παρατηρήσεων O, και το μοντέλο λ. Η εξίσωση (3.14) μπορεί να εκφραστεί απλά με βάση τις forward και backward μεταβλητές δηλαδή,

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)} \tag{3.15}$$

Ο παράγοντας κανονικοποίησης P(O|λ) κάνει το γ ένα μέτρο πιθανότητας έτσι ώστε αθροίζοντας για όλα τα γt(i) επί του N να λαμβάνουμε τη μονάδα. Χρησιμοποιώντας το γ μπορούμε να επιλύσουμε για το μεμονωμένα πλέον πιθανό q τη χρονική στιγμή t ως

$$q_t = \underset{1 \leqslant i \leqslant N}{argmax}\left[\gamma_t(i)\right] \qquad 1 \leqslant t \leqslant T \tag{3.16}$$

Αν και η εξίσωση (3.16) μεγιστοποιεί τον αναμενόμενο αριθμό σωστών καταστάσεων (με την επιλογή της πλέον πιθανής κατάστασης για κάθε t), θα μπορούσαν να υπάρξουν μερικά προβλήματα με την προκύπτουσα ακολουθία καταστάσεων. Παραδείγματος χάριν, όταν το HMM έχει μεταβάσεις καταστάσεων με μηδενική πιθανότητα, η βέλτιστη ακολουθία μπορεί στην πραγματικότητα ακόμα και να μην είναι μια έγκυρη ακολουθία καταστάσεων. Αυτό οφείλεται στο γεγονός ότι η λύση της (3.16) απλά καθορίζει την πλέον πιθανή κατάσταση σε κάθε στιγμή, αδιαφορώντας για την πιθανότητα εμφάνισης ακολουθιών καταστάσεων.

Μια πιθανή λύση στο ανωτέρω πρόβλημα είναι να τροποποιηθεί το κριτήριο βελτιστότητας. Παραδείγματος χάριν, θα μπορούσαμε να επιλύσουμε το πρόβλημα για την

ακολουθία καταστάσεων που μεγιστοποιεί τον αναμενόμενο αριθμό σωστών ζευγαριών καταστάσεων ή τριάδων καταστάσεων. Αν και αυτά τα κριτήρια μπορεί να είναι λογικά για μερικές εφαρμογές, το ευρύτερα χρησιμοποιούμενο κριτήριο είναι να βρίσκεται η μεμονωμένη καλύτερη κατάσταση κάθε φορά, δηλ., για να μεγιστοποιήσουμε την πιθανότητα P(Q|O, λ) που είναι ισοδύναμη με τη μεγιστοποίηση της P(Q, O|λ). Μια τυποποιημένη τεχνική για αυτήν την μεμονωμένη καλύτερη κατάσταση είναι ο αλγόριθμος Viterbi.

Αλγόριθμος Viterbi [21], [22]: Για να βρεθεί η μεμονωμένα καλύτερη ακολουθία καταστάσεων , Q = {ql q2… qT}, για τη δεδομένη ακολουθία O παρατηρήσεων = (O1, O2,...OT}, πρέπει να ορίσουμε την ποσότητα

$$\delta_t(i) = \max_{q_1, q_2 \ldots q_{t-1}} P\left[q_1 q_2 \ldots q_t = i, O_1 O_2 \ldots O_t | \lambda\right] \tag{3.17}$$

όπου με δ συμβολίζουμε το καλύτερο αποτέλεσμα (υψηλότερη πιθανότητα) κατά μήκος ενός μονοπατιού τη χρονική στιγμή t, που αφορά τις πρώτες t παρατηρήσεις και τερματίζει στην κατάσταση Si. Αναδρομικά

$$\delta_{t+1}(j) = \left[\max_i \delta_t(i) a_{ij}\right] \cdot b_j(O_{t+1}) \tag{3.18}$$

Για να ανακτήσουμε πραγματικά την ακολουθία καταστάσεων, πρέπει να παρακολουθήσουμε το όρισμα που μεγιστοποίησε την (3.18) για κάθε t και j. Αυτό επιτυγχάνεται μέσω του πίνακα ψt(j). Η πλήρης διαδικασία για την εύρεση της καλύτερης ακολουθίας καταστάσεων μπορεί τώρα να οριστεί ως εξής:

1) Αρχικοποίηση:

$$\delta_1(j) = \pi_i b i(O_1) \quad 1 \leqslant i \leqslant N \tag{3.19α}$$

$$\psi_1(i) = 0$$

2) Αναδρομή:

$$\delta_t(j) = \max_{1 \leqslant i \leqslant N}\left[\delta_{t-1}(i) a_{ij}\right] b_j(O_t) \quad 2 \leqslant t \leqslant T, \ 1 \leqslant j \leqslant N \tag{3.19β}$$

$$\psi_t(j) = \operatorname*{argmax}_{1 \leqslant i \leqslant N}\left[\delta_{t-1}(i) a_{ij}\right] \quad 2 \leqslant t \leqslant T, \ 1 \leqslant j \leqslant N$$

3)Τερματισμός

$$p* = \max_{1 \leqslant i \leqslant N}\left[\delta_T(i)\right] \tag{3.19γ}$$

$$q*_T = \operatorname*{argmax}_{1 \leqslant i \leqslant N}\left[\delta_T(i)\right]$$

4) Οπισθοδρόμηση:

$$q*_t = \psi_{t+1}(q_{t+1}) \tag{3.19δ}$$

Πρέπει να σημειωθεί ότι ο αλγόριθμος του Viterbi είναι παρόμοιος (εκτός από το βήμα της οπισθοδρόμησης) στην υλοποίηση με τον forward αλγόριθμο. Η μεγαλύτερη διαφορά είναι η μεγιστοποίηση της (3.19β) επί των προηγουμένων καταστάσεων που χρησιμοποιείται αντί της διαδικασίας αθροίσματος της (3.13β). Πρέπει επίσης να γίνει σαφές ότι μια δομή πλέγματος μπορεί να υλοποιήσει αποτελεσματικά τον υπολογισμό της διαδικασίας Viterbi.

3.4.3. Πρόβλημα 3

Το τρίτο, και κατά πολύ δυσκολότερο, πρόβλημα των HMMs [2]-[6] είναι να προσδιοριστεί μια μέθοδος για να ρυθμιστούν οι παράμετροι του μοντέλου (A, B, π) ώστε να μεγιστοποιήσουμε την πιθανότητα της ακολουθίας παρατηρήσεων δοσμένου του μοντέλου. Δεν υπάρχει κανένας γνωστός τρόπος αναλυτικής επίλυσης για το μοντέλο που μεγιστοποιεί την πιθανότητα της ακολουθίας παρατηρήσεων. Στην πραγματικότητα, δοθείσης μιας πεπερασμένης ακολουθίας παρατηρήσεων ως δεδομένο εκπαίδευσης, δεν υπάρχει κανένας βέλτιστος τρόπος υπολογισμού των παραμέτρων του μοντέλου. Μπορούμε, εντούτοις, να επιλέξουμε το λ = (A, B, π) έτσι ώστε η P(O|λ) να μεγιστοποιείται τοπικά χρησιμοποιώντας μιας επαναληπτική διαδικασία όπως ο αλγόριθμος Baum-Welch (EM method)[23] ή χρησιμοποιώντας τεχνικές κλίσης [15]. Σε αυτή την ενότητα παρουσιάζεται μια επαναληπτική διαδικασία, βασισμένη πρώτιστα στην κλασική εργασία του Baum και των συναδέλφων του, για την επιλογή των παραμέτρων του μοντέλου.

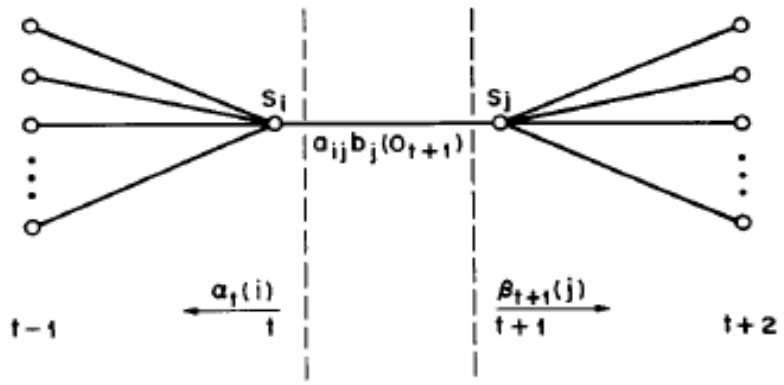
Προκειμένου να περιγραφεί η διαδικασία για τον επανυπολογισμό (επαναληπτική ανανέωση και βελτίωση) των παραμέτρων του HMM, αρχικά ορίζουμε το ξ(i,j), την πιθανότητα το σύστημα να βρίσκεται στην κατάσταση Si τη χρονική στιγμή t, και στην κατάσταση Sj τη χρονική στιγμή t+1, δοθέντος του μοντέλου και της ακολουθίας παρατηρήσεων

$$\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j | O, \lambda) \tag{3.20}$$

Η ακολουθία γεγονότων που οδηγεί στις συνθήκες που απαιτούνται από την (3.20) επιδεικνύονται στην εικόνα 3. Είναι σαφές, από τους ορισμούς των forward και backward μεταβλητών, ότι μπορούμε να γράψουμε το ξ(i,j) στη μορφή

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O|\lambda)} \tag{3.21}$$

όπου ο αριθμητής είναι ακριβώς P(qt = Si, qt+l = Sj, O|λ)και η διαίρεση με το P(O|λ) δίνει το επιθυμητό μέτρο πιθανότητας.



Εικόνα 3.4 Απεικόνιση της ακολουθίας ενεργειών που απαιτούνται για τον υπολογισμό του γεγονότος ότι το σύστημα είναι στην κατάσταση Si τη χρονική στιγμή t και στην κατάσταση Sj τη χρονική στιγμή t+1

Έχουμε ορίσει προηγουμένως το γt(i) ως την πιθανότητα εύρεσης του συστήματος

στην κατάσταση Si τη χρονική στιγμή t, δοθείσης της ακολουθίας παρατηρήσεων και του μοντέλου ως εκ τούτου μπορούμε να συσχετίσουμε τα γ και ξ αθροίζοντας επί του j, λαμβάνοντας

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j) \qquad \textbf{(3.22)}$$

Αθροίζοντας επί του χρόνου t λαμβάνουμε μια ποσότητα που μπορεί να ερμηνευθεί ως ο αναμενόμενος (με την πάροδο του χρόνου) αριθμός φορών που επισκεπτόμαστε την κατάσταση Si , ή ισοδύναμα, ο αναμενόμενος αριθμός μεταβάσεων που γίνονται από την κατάσταση Si (εάν αποκλείουμε τη χρονική στιγμή t = T από το άθροισμα). Ομοίως, το άθροισμα των ξ(i, j) επί του t(από t = 1 ως t=T-1) μπορεί να ερμηνευθεί ως ο αναμενόμενος αριθμός μεταβάσεων από την κατάσταση Si στην Sj. Ήτοι,

$$\sum_{t=1}^{T-1} \gamma_t(i) = \textbf{αναμενόμενος αριθμός μεταβάσεων από το Si}$$

$$\sum_{t=1}^{T-1} \xi_t(i,j) = \textbf{αναμενόμενος αριθμός μεταβάσεων από το Si στο Sj}$$

Χρησιμοποιώντας τους ανωτέρω τύπους (και την έννοια της μέτρησης των εμφανίσεων των ανωτέρω γεγονότων) μπορούμε να δώσουμε μια μέθοδο για τον επανυπολογισμό των παραμέτρων ενός HMM. Ένα σύνολο τέτοιων τύπων επανυπολογισμού για τα π, A, B είναι

$$\overline{\pi_i} = \textbf{αναμενόμενη συχνότητα στην κατάσταση Si τη χρονική στιγμή (t=1)= } \gamma_1(i) \qquad \textbf{(3.23α)}$$

$$\overline{a_{ij}} = \frac{\textbf{αναμενόμενος αριθμός μεταβάσεων από την κατάσταση } S_i \textbf{ στην κατάσταση } S_j}{\textbf{αναμενόμενος αριθμός μεταβάσεων από την κατάσταση } S_i}$$

$$\overline{a_{ij}} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \qquad \textbf{(3.23β)}$$

$$\overline{b_i}(k) = \frac{\textbf{αναμενόμενος αριθμός φορών στην κατάσταση j παρατηρώντας το σύμβολο } v_k}{\textbf{αναμενόμενος αριθμός φορών στην κατάσταση j}}$$

$$\overline{b_i}(k) = \frac{\sum_{t=1, O_t=v_k}^{T-1} \gamma_t(j)}{\sum_{t=1}^{T-1} \gamma_t(j)} \qquad \textbf{(3.23γ)}$$

Εάν ορίσουμε το παρόν μοντέλο ως λ = (A, B, π), και το χρησιμοποιήσουμε για να υπολογίσουμε τα δεξιά τμήματα των εξισώσεων (3.23) και στη συνέχεια ορίσουμε το νέο μοντέλο από τις αριστερές πλευρές των (3.23), έχει αποδειχθεί από τον Baum και τους

συναδέλφους του [7], [4] ότι είτε 1) το αρχικό μοντέλο λ ορίζει ένα κρίσιμο σημείο της συνάρτησης πιθανότητας, οπότε και ισχύει $\bar{\lambda}$ =λ, είτε 2) μοντέλο $\bar{\lambda}$ είναι πιθανότερο από το μοντέλο λ υπό την έννοια ότι $P(O|\bar{\lambda})>P(O|\lambda)$ δηλαδή έχουμε βρει ένα νέο μοντέλο από το οποίο είναι πιο πιθανό να έχει παραχθεί η ακολουθία παρατηρήσεων.

Με βάση την ανωτέρω διαδικασία, εάν χρησιμοποιήσουμε το καινούριο μοντέλο που προκύπτει κάθε φορά μετά τον υπολογισμό των (3.23) και επαναλάβουμε αυτή τη διαδικασία μπορούμε να βελτιώσουμε την πιθανότητα εμφάνισης της O από το μοντέλο έως ότου επιτύχουμε κάποιο όριο. Το τελικό αποτέλεσμα αυτής της διαδικασίας καλείται maximum likelihood estimate (υπολογισμός μέγιστης πιθανότητας) του HMM. Πρέπει να επισημανθεί ότι ο forward-backward αλγόριθμος οδηγεί μόνο σε τοπικά μέγιστα, και στα περισσότερα προβλήματα που μας ενδιαφέρουν, η επιφάνεια βελτιστοποίησης είναι πολύ σύνθετη και έχει πολλά τοπικά μέγιστα.

Οι τύποι επανυπολογισμού (3.23α)-(3.23γ) μπορούν να παραχθούν άμεσα μεγιστοποιώντας τη βοηθητική συνάρτηση Baum

$$Q(\lambda,\bar{\lambda}) = \sum_{Q} P(Q|O,\lambda)\left[\log P(O,Q|\bar{\lambda})\right] \tag{3.24}$$

επί του $\bar{\lambda}$. Έχει αποδειχθεί από τον Baum και τους συναδέλφους του [7], [4] ότι η μεγιστοποίηση της Q οδηγεί σε αυξημένη πιθανότητα. Τελικά η συνάρτηση πιθανότητας συγκλίνει σε ένα κρίσιμο σημείο.

Σχόλιο. Οι τύποι επανυπολογισμού μπορούν εύκολα να ερμηνευθούν ως υλοποίηση του αλγορίθμου EM της στατιστικής [23] στον οποίο το βήμα E είναι ο υπολογισμός της βοηθητικής συνάρτησης Q, και το βήμα M (τροποποίηση) είναι η μεγιστοποίηση επί του λ. Κατά συνέπεια οι εξισώσεις Baum-Welch είναι ουσιαστικά ίδιες με τα βήματα του EM για το συγκεκριμένο πρόβλημα.

Μια σημαντική πτυχή της reestimation διαδικασίας είναι ότι οι στοχαστικοί περιορισμοί των παραμέτρων του HMM, δηλαδή

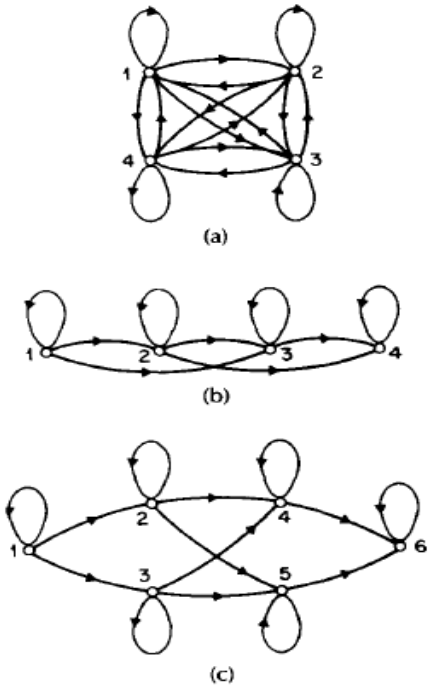
$$\sum_{i=1}^{N} \bar{\pi}_i = 1$$

$$\sum_{j=1}^{N} \bar{a}_{ij} = 1 \quad 1 \leqslant i \leqslant N$$

$$\sum_{k=1}^{M} \bar{b}_j(k) = 1 \quad 1 \leqslant j \leqslant N$$

ικανοποιούνται αυτόματα σε κάθε επανάληψη. Τέλος, σημειώνουμε ότι αφού ολόκληρο το πρόβλημα μπορεί να θεωρηθεί ως πρόβλημα βελτιστοποίησης, οι τυποποιημένες τεχνικές κλίσης μπορούν να χρησιμοποιηθούν για να επιλύσουν τις «βέλτιστες» τιμές των μοντέλων παραμέτρων [15]. Τέτοιες διαδικασίες έχουν δοκιμαστεί και έχουν αποδειχθεί να παράγουν λύσεις συγκρίσιμες με αυτές των τυποποιημένων διαδικασιών επανυπολογισμού (reestimation).

3.5 Τύποι HMMs



Εικόνα 3.5 Παρουσίαση 3 τύπων HMM: a)4-state ergodic model b) 4-state left-right model c) 6-state parallel path model

Μέχρι τώρα, έχουμε εξετάσει μόνο την ειδική περίπτωση εργοδικών ή πλήρως συνδεδεμένων HMMs στα οποίο κάθε κατάσταση του μοντέλου θα μπορούσε να επιτευχθεί (σε ένα μόνο βήμα) από κάθε άλλη κατάσταση του μοντέλου. (τυπικά, εργοδικό καλείται το μοντέλο που έχει την ιδιότητα κάθε κατάστασή του να μπορεί να επιτευχθεί από κάθε άλλη κατάσταση σε έναν πεπερασμένο αριθμό βημάτων).

Για κάποιες εφαρμογές, άλλοι τύποι HMMs έχουν βρεθεί να περιγράφουν τις παρατηρηθείσες ιδιότητες του σήματος καλύτερα από το τυπικό εργοδικό μοντέλο. Ένα τέτοιο μοντέλο παρουσιάζεται στην εικόνα 3.5b. Αυτό το μοντέλο καλείται left-right μοντέλο ή μοντέλο Bakis επειδή η ακολουθία καταστάσεων του μοντέλου έχει την ιδιότητα ότι με την πάροδο του χρόνου ο δείκτης καταστάσεων αυξάνει (ή μένει ο ίδιος), δηλαδή, οι καταστάσεις προχωρούν από τα αριστερά στα δεξιά. Σαφώς ο τύπος αυτός HMM έχει την ιδιότητα ότι μπορεί εύκολα να μοντελοποιήσει σήματα των οποίων ιδιότητες αλλάζουν με το πέρασμα του χρόνου (πχ ομιλία). Η θεμελιώδης ιδιότητα όλων των left-right HMM είναι ότι οι συντελεστές μετάβασης καταστάσεων έχουν την ιδιότητα aij = 0, j < i δηλαδή, καμία μετάβαση δεν επιτρέπεται στις καταστάσεις των οποίων οι δείκτες είναι χαμηλότεροι από τη τρέχουσα κατάσταση. Επιπλέον, οι πιθανότητες της αρχικής κατάστασης έχουν την ιδιότητα

$$\pi_i = \begin{matrix} 0, & i \neq 1 \\ 1, & i = 1 \end{matrix}$$

αφού η ακολουθία καταστάσεων πρέπει να αρχίσει στην κατάσταση 1 (και να τελειώσει στην κατάσταση N). Συχνά, με τα left-right μοντέλα, πρόσθετοι περιορισμοί επιβάλλονται στους συντελεστές μετάβασης καταστάσεων ώστε να εξασφαλιστεί ότι δεν θα εμφανιστούν μεγάλες αλλαγές στους δείκτες καταστάσεων. Ως εκ τούτου συχνά χρησιμοποιείται ένας περιορισμός της μορφής

aij=0, j > i + Δ.

Είναι φανερό, ότι για την τελευταία κατάσταση σε ένα left-right μοντέλο, οι συντελεστές

μετάβασης καταστάσεων έχουν τις τιμές
$\mathbf{a_{NN} = 1}$
$\mathbf{a_{Ni} = 0, i<N}$

Αν και έχουμε διαιρέσει τα HMMs στα εργοδικά και τα left-right μοντέλα, προφανώς υπάρχουν πολλές πιθανές παραλλαγές και συνδυασμοί. Για παράδειγμα, στην εικόνα 3.5c φαίνεται ένα parallel path left-right model που αν και τυπικά αποτελεί ένα left-right μοντέλο εντούτοις, είναι φανερό πως έχει κάποια ορισμένη ευελιξία μη παρούσα σε ένα τυπικό left-right μοντέλο (δηλαδή, ένα χωρίς παράλληλες πορείες).

Πρέπει να είναι επίσης σαφές ότι η επιβολή των περιορισμών των left-right μοντέλων , ή εκείνων των περιορισμένων αλμάτων, δεν έχει ουσιαστικά καμία επίδραση στη διαδικασία επανυπολογισμού. Αυτό συμβαίνει επειδή οποιαδήποτε παράμετρος του HMM που τίθεται μηδέν αρχικά, θα παραμείνει μηδέν σε όλη τη διαδικασία reestimation.

Συνεχείς πυκνότητες παρατήρησης σε HMMs

Μέχρι στιγμής έχουμε εξετάσει μόνο την περίπτωση όπου οι παρατηρήσεις αποτελούσαν διακριτά σύμβολα από ένα πεπερασμένο αλφάβητο και γι' αυτό μπορούσαμε να χρησιμοποιήσουμε μια διακριτή πυκνότητα πιθανότητας μέσα σε κάθε κατάσταση αυτού του μοντέλου. Το πρόβλημα με αυτήν την προσέγγιση, τουλάχιστον για μερικές εφαρμογές, είναι ότι οι παρατηρήσεις είναι συνεχή σήματα (ή διανύσματα). Αν και είναι δυνατό να κβαντισθούν τέτοια συνεχή σήματα μέσω σημάτων αναφοράς μπορεί να υπάρξει σοβαρή υποβάθμιση λόγω αυτής. Ως εκ τούτου θα ήταν προτιμότερο να είμαστε σε θέση να χρησιμοποιήσουμε HMMs με συνεχείς πυκνότητες παρατήρησης.

Προκειμένου να χρησιμοποιηθεί μια συνεχής πυκνότητα παρατήρησης, πρέπει να τοποθετηθούν κάποιοι περιορισμοί στη μορφή της συνάρτησης πυκνότητας πιθανότητας (pdf) του μοντέλου για να εξασφαλιστεί ότι οι παράμετροι της σππ μπορούν να επανυπολογισθούν με έναν συνεπή τρόπο. Η γενικότερη μορφή της σππ, για την οποία μια διαδικασία επανυπολογισμού έχει διατυπωθεί [24] - [26], είναι μια πεπερασμένη μίξη της μορφής

$$b_j(O) = \sum_{m=1}^{M} c_{jm} \Pi\left[O, \mu_{jm}, U_{jm}\right], \quad 1 \leqslant j \leqslant N \tag{3.25}$$

όπου O είναι το διάνυσμα που μοντελοποιείται, cjm είναι ο συντελεστής μίξης για την m-στη μίξη στην κατάσταση j και Π είναι οποιαδήποτε λογαριθμικά κοίλη ή ελλειπτική συμμετρική πυκνότητα [24] (π.χ., γκαουσιανή), με μέσο διάνυσμα μjm και μήτρα συνδιακύμανσης Ujm, για τον m-οστό όρο μίξης στη κατάσταση j. Συνήθως χρησιμοποιείται για το Π μια γκαουσιανή πυκνότητα. Τα κέρδη των cjm, ικανοποιούν τον στοχαστικό περιορισμό

$$\begin{aligned} &\sum_{m=1}^{M} c_{jm} = 1 \quad 1 \leqslant j \leqslant N \\ &c_{jm} \geqslant 0 \quad 1 \leqslant j \leqslant N, \ 1 \leqslant m \leqslant M \end{aligned} \tag{3.26}$$

έτσι ώστε η σππ να είναι κατάλληλα κανονικοποιημένη, δηλαδή,

$$\int_{-\infty}^{+\infty} b_j(x) d(x) = 1, \quad 1 \leqslant j \leqslant N \tag{3.27}$$

Η συνάρτηση πυκνότητας πιθανότητας της εξίσωσης (3.25) μπορεί να χρησιμοποιηθεί για να προσεγγίσει, με αυθαίρετη ακρίβεια, οποιαδήποτε πεπερασμένη, συνεχή συνάρτηση πυκνότητας. Ως εκ τούτου μπορεί να εφαρμοστεί σε ένα ευρύ φάσμα προβλημάτων.

Μπορεί να αποδειχθεί [24] - [26] ότι οι τύποι reestimation για τους συντελεστές της πυκνότητας των cjm είναι της μορφής

$$\overline{c_{jk}} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)}{\sum_{t=1}^{T}\sum_{k=1}^{M} \gamma_t(j,k)} \tag{3.27}$$

$$\overline{\mu_{jk}} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)\cdot O_t}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{3.28}$$

$$\overline{U_{jk}} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)\cdot(O_t-\mu_{jk})(O_t-\mu_{jk})'}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{3.29}$$

όπου η απόστροφος συμβολίζει το ανάστροφο διάνυσμα και γt(j, k) είναι η πιθανότητα το σύστημα να βρίσκεται στην κατάσταση j τη χρονική στιγμή t με τον k-οστό όρο μίξης για την ακολουθία O.

$$\gamma_t(j,k) = \frac{\alpha_t(j)\beta_t(j)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} \cdot \frac{c_{jk}\Pi(O,\mu_{jk},U_{jk})}{\sum_{m=1}^{M} c_{jm}\Pi(O,\mu_{jm},U_{jm})}$$

Ο τύπος reestimation για το aij είναι ίδιος με αυτόν που χρησιμοποιείται για τις διακριτές πυκνότητες παρατήρησης. Επίσης ο τύπος reestimation για το cjk είναι η αναλογία μεταξύ των αναμενόμενων φορών που το σύστημα είναι στην κατάσταση j χρησιμοποιώντας τον k-οστό όρο μίξης. Ομοίως, ο τύπος reestimation για το μέσο διάνυσμα μjk πολλαπλασιάζει κάθε αριθμητή της (3.27) με την παρατήρηση, δίνοντας έτσι την αναμενόμενη τιμή του τμήματος του διανύσματος παρατηρήσεων που αποτελείται από τον k-οστό όρο μίξης. Μια παρόμοια ερμηνεία μπορεί να δοθεί για τον όρο reestimation για τη μήτρα συνδιακύμανσης Ujk.

Αυτο-αναδρομικά HMMs

Αν και η γενική διατύπωση των συνεχούς πυκνότητας HMMs ισχύει σε έναν ευρύ φάσμα των προβλημάτων, υπάρχει μια άλλη πολύ ενδιαφέρουσα κατηγορία HMMs που

μπορεί να εφαρμοστεί ιδιαίτερα στην επεξεργασία λόγου. Αυτό είναι η κατηγορία των αυτο-αναδρομικών HMMs [27], [28]. Για αυτήν την κατηγορία, τα διανύσματα παρατηρήσεων προέρχονται από μια αυτο-αναδρομική διαδικασία.

Πιο συγκεκριμένα, θεωρούμε το διάνυσμα παρατηρήσεων O με στοιχεία (x0, x1, x2, …, xk-1). Αφού η συνάρτηση πυκνότητας πιθανότητας βάσης για το διάνυσμα παρατηρήσεων είναι γκαουσιανή αυτο-αναδρομική (ή τάξης p), τα στοιχεία του O συσχετίζονται ως εξής:

$$O_k = -\sum_{i=1}^{p} a_i O_{k-i} + e_k \tag{3.30a}$$

όπου ek, k = 0,1,2,…, k - 1 είναι γκαουσσιανές, ανεξάρτητες, όμοια κατανεμημένες τυχαίες μεταβλητές με μηδενική μέση τιμή και διακύμανση σ^2 και ai, i = 1,2,...,p είναι οι συντελεστές αυτο-αναδρομής. Αποδεικνύεται ότι για μεγάλες τιμές του k, η συνάρτηση πυκνότητας για το O είναι προσεγγιστικά

$$f(O) = (2\pi\sigma^2)^{-K/2} \exp\left(-\frac{1}{2\sigma^2}\delta(O,a)\right) \tag{3.30b}$$

όπου

$$\delta(O,a) = r_a(0)r(0) + 2\sum_{i=1}^{p} r_a(i)r(i) \tag{3.30c}$$

$$a' = \left[1, a_1, a_2, \ldots a_p\right] \tag{3.30d}$$

$$r_a(i) = \sum_{n=0}^{p-i} a_n a_{n+i} \quad (a_0 = 1), 1 \leqslant i \leqslant p \tag{3.30e}$$

$$r(i) = \sum_{n=0}^{K-i-1} x_n x_{n+i} \quad 0 \leqslant i \leqslant p \tag{3.30f}$$

Στις ανωτέρω εξισώσεις είναι φανερό ότι το r(i) είναι η αυτοσυσχέτιση των δειγμάτων παρατηρήσεων, και το ra(i) είναι η αυτοσυσχέτιση των αυτο-αναδρομικών συντελεστών.
Το συνολικό υπόλοιπο πρόβλεψης μπορεί να γραφτεί ως

$$\alpha = E\left[\sum_{i=1}^{K} (e_i)^2\right] = K\sigma^2 \tag{3.31}$$

όπου σ^2 είναι η διασπορά ανά δείγμα του σήματος σφάλματος. Έστω το κανονικοποιημένο διάνυσμα παρατηρήσεων

$$\hat{O} = \frac{O}{\sqrt{\alpha}} = \frac{O}{\sqrt{K\sigma^2}} \tag{3.32}$$

όπου κάθε δείγμα xi διαιρείται με το $\sqrt{K\sigma^2}$,δηλαδή κάθε δείγμα είναι κανονικοποιημένο από τη διασπορά των δειγμάτων. Κατόπιν η f(O) μπορεί να γραφτεί ως

$$f(\hat{O}) = \left(\frac{2\pi}{K}\right)^{-K/2} \exp\left(-\frac{K}{2}\,\delta(\hat{O}, a)\right) \tag{3.33}$$

Στην πράξη, ο παράγοντας K μπροστά από τον εκθέτη αντικαθίσταται από ένα αποτελεσματικό μήκος K πλαισίων που αντιπροσωπεύει το πραγματικό μήκος κάθε διανύσματος δεδομένων. Κατά συνέπεια, εάν τα διαδοχικά διανύσματα δεδομένων επικαλύπτονται (3 σε 1) θα πρέπει να χρησιμοποιήσουμε K = K/3 στην (3.33), έτσι ώστε η συμβολή κάθε δείγματος του σήματος στη γενική πυκνότητα να λαμβάνεται υπόψιν ακριβώς μια φορά.

Ο τρόπος με τον οποίο χρησιμοποιούμε τη γκαουσιανή αυτο-αναδρομική πυκνότητα στα HMMs είναι απλός. Υποθέτουμε μια πυκνότητα μίξης της μορφής

$$b_j(O) = \sum_{m=1}^{M} (c_{jm})\, b_{jm(O)} \tag{3.34}$$

όπου κάθε bjm είναι η πυκνότητα που καθορίζεται από την (3.33) με διάνυσμα αυτο-αναδρομής a = ajm

Ένας τύπος reestimation για την ακολουθία αυτοσυσχέτισης, r(i), για το στοιχείο κατάστασης j και k-οστής μίξης, έχει την μορφή

$$\bar{r}_{jk} = \frac{\sum_{t=1}^{T} (\gamma_t)(j,k)\, r_t}{\sum_{t=1}^{T} (\gamma_t)(j,k)} \tag{3.35}$$

όπου η ποσότητα γt(j,k) ορίζεται ως η πιθανότητα να βρίσκεται το σύστημα στην κατάσταση i τη χρονική στιγμή t και χρησιμοποιώντας το τμήμα μίξης k, ήτοι

$$\gamma_t(j,k) = \left[\frac{\alpha_t(j)\beta_t(j)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)}\right]\left[\frac{c_{jk}(jk)\, b_{jk}(O_t)}{\sum_{k=1}^{M} c_{jk}(jk)\, b_{jk}(O_t)}\right] \tag{3.36}$$

Γίνεται φανερό ότι η ποσότητα rjk είναι ένα σταθμισμένο άθροισμα (από την πιθανότητα της εμφάνισης) των κανονικοποιημένων αυτοσυσχετίσεων των πλαισίων στην ακολουθία παρατηρήσεων. Από το rjk, μπορεί κανείς να λύσει ένα σύνολο κανονικών εξισώσεων για να λάβει το αντίστοιχο αυτοαναδρομικό διάνυσμα ajk, για το k-οστό μίγμα της κατάστασης j. Τα νέα διανύσματα αυτο-διόρθωσης των συντελεστών αυτοαναδρομής μπορούν στη συνέχεια να υπολογιστούν από την (3.30e).

Κριτήριο βελτιστοποίησης - MML, MMI, και MDI

Η βασική φιλοσοφία των HMMs είναι ότι ένα σήμα (ή μια ακολουθία παρατηρήσεων) μπορεί να μοντελοποιηθεί καλά εάν οι παράμετροι του HMM επιλεχθούν προσεκτικά και ορθά. Το πρόβλημα με αυτήν την φιλοσοφία είναι ότι είναι μερικές φορές ανακριβής είτε

επειδή το σήμα δεν υπακούει τους περιορισμούς του HMM, είτε επειδή είναι πάρα πολύ δύσκολο να αποκτηθούν αξιόπιστες εκτιμήσεις όλων των παραμέτρων του HMM. Για να υπερκεραστεί αυτός ο τύπος προβλήματος, έχουν προταθεί τουλάχιστον δύο εναλλακτικές λύσεις στην τυπική διαδικασία βελτιστοποίησης maximum likelihood (μέγιστης πιθανότητας) ML για τις παραμέτρους του HMM.

Η πρώτη εναλλακτική λύση [30] είναι βασισμένη στην ιδέα ότι πρόκειται να σχεδιαστούν περισσότερα του ενός HMM και επιθυμούμε να τα σχεδιάσουμε όλα συγχρόνως με τέτοιο τρόπο ώστε να μεγιστοποιηθεί η δυνατότητα διάκρισης κάθε μοντέλου (δηλαδή, η δυνατότητα κάθε μοντέλου να διακρίνει μεταξύ των ακολουθιών παρατηρήσεων που παράγονται από το σωστό μοντέλο και εκείνων που παράγονται από τα εναλλακτικά μοντέλα). Συμβολίζουμε τα διαφορετικά HMMs ως λ_v, $v = 1, 2, \ldots, V$. Το τυπικό κριτήριο ML είναι να χρησιμοποιηθεί μια διαφορετική ακολουθία παρατηρήσεων εκπαίδευσης O^v για να παράγει τις παραμέτρους του μοντέλου για κάθε μοντέλο λ. Κατά συνέπεια, η τυπική βελτιστοποίηση ML θα είναι

$$P_v* = \max_{\lambda_v} P(O|\lambda_v) \tag{3.37}$$

Το προτεινόμενο εναλλακτικό κριτήριο σχεδίασης [29] είναι το κριτήριο μέγιστης αμοιβαίας πληροφορίας (MMI) στο οποίο η μέσα αμοιβαία πληροφορία I μεταξύ της ακολουθίας παρατηρήσεων O και του συνόλου των μοντέλων λ = (λ1,λ2,...λν) μεγιστοποιείται. Ένας δυνατός τρόπος υλοποίησης του παραπάνω είναι

$$I_v* = \max_{\lambda}\left[\log P(O|\lambda_v) - \log \sum_{w=1}^{V} P(O|\lambda_w)\right] \tag{3.38}$$

δηλαδή, να επιλέξουμε το λ ώστε να χωριστεί το σωστό μοντέλο λν, από όλα τα άλλα μοντέλα στην ακολουθία εκπαίδευσης O^v. Αθροίζοντας στην (3.38) για όλες τις ακολουθίες εκπαίδευσης, ελπίζουμε να επιτύχουμε το πλέον χωρισμένο σύνολο μοντέλων που είναι δυνατόν. Κατά συνέπεια μια πιθανή υλοποίησης θα ήταν

$$I* = \max_{\lambda}\{\sum_{v=1}^{V}\left[\log P(O^v|\lambda_v) - \log \sum_{w=1}^{V} P(O^v|\lambda_w)\right]\} \tag{3.39}$$

Υπάρχουν διάφοροι θεωρητικοί λόγοι για τους οποίους δεν μπορούν να πραγματοποιηθούν αναλυτικές (ή με τύπο reestimation) λύσεις για την (3.39). Κατά συνέπεια, ο μόνος γνωστός τρόπος επίλυσης της (3.39) είναι μέσω γενικών μεθόδων βελτιστοποίησης όπως οι μέθοδοι απότομης καθόδου.

Η δεύτερη εναλλακτική φιλοσοφία είναι να υποθέσουμε ότι το σήμα που μοντελοποιείται δεν παρήχθη απαραιτήτως από μια πηγή Markov, αλλά υπακούει ορισμένους περιορισμούς (π.χ., θετική καθορισμένη συνάρτηση συσχετισμού) [31]. Ο στόχος της διαδικασίας σχεδιασμού είναι επομένως να επιλεχθούν οι παράμετροι του HMM που ελαχιστοποιούν την πληροφορία διάκρισης (DI) ή την εντροπία (cross entropy) μεταξύ του συνόλου των έγκυρων (δηλαδή αυτών που ικανοποιούν τις μετρήσεις) πυκνοτήτων πιθανοτήτων των σημάτων (έστω ότι το σύνολο αυτό συμβολίζεται με Q) και του σύνολου πυκνοτήτων πιθανοτήτων των HMM (έστω ότι το σύνολο αυτό συμβολίζεται με Pλ), όπου η DI μεταξύ των Q και Pλ μπορεί να γραφεί γενικά στη μορφή

$$D(Q\|P_\lambda)=\int q(y)\ln(q(y)/p(y))dy \qquad (3.40)$$

όπου τα q και p είναι οι συναρτήσεις πυκνότητας πιθανότητας που αντιστοιχούν στα Q και Pλ. Οι υπάρχουσες τεχνικές ελαχιστοποίησης της (3.40) (με αυτόν τον τρόπο δίνοντας μια λύση MDI) για τις βέλτιστες τιμές λ =(A, B, π) είναι ιδιαίτερα πολύπλοκες εντούτοις, χρησιμοποιούν έναν γενικευμένο αλγόριθμο Baum ως πυρήνα κάθε επανάληψης και έτσι προσαρμόζονται αποτελεσματικά στα HMM.[31]

Έχει αποδειχθεί ότι οι προσεγγίσεις ML, MMI, και MDI μπορούν να διατυπωθούν ομοιόμορφα ως προσεγγίσεις MDI. Οι τρεις προσεγγίσεις διαφέρουν είτε στην πυκνότητα πιθανότητας που αποδίδεται στην πηγή που μοντελοποιείται, είτε στο μοντέλο που χρησιμοποιείται. Καμία από τις προσεγγίσεις, εντούτοις, δεν υποθέτει ότι η πηγή έχει τη κατανομή πιθανότητας του μοντέλου.

3.6 Σύγκριση HMMs - Ένα πιθανοτικό μέτρο απόστασης για HMMs

Μια ενδιαφέρουσα ερώτηση που συνδέεται με HMMs είναι η ακόλουθη: δοθέντων δύο HMMs, *λ1* και *λ*, ποιο είναι ένα λογικό μέτρο της ομοιότητας των δύο μοντέλων; Ένα σημείο κλειδί εδώ είναι το κριτήριο ομοιότητας. Έστω η περίπτωση δύο μοντέλων

$$\lambda_1 = (A_1, B_1, \pi_1) \quad \lambda_2 = (A_2, B_2, \pi_2)$$

με

$$A_1 = \begin{matrix} p & 1-p \\ 1-p & p \end{matrix} \quad B_1 = \begin{matrix} q & 1-q \\ 1-q & q \end{matrix} \quad \pi_1 = 1/2 \quad 1/2$$

$$A_2 = \begin{matrix} r & 1-r \\ 1-r & r \end{matrix} \quad B_2 = \begin{matrix} s & 1-s \\ 1-s & s \end{matrix} \quad \pi_2 = 1/2 \quad 1/2$$

Για να είναι το *λ1* ισοδύναμο με το *λ2*, με την έννοια ότι θα έχουν ίδιες στατιστικές ιδιότητες για τα σύμβολα παρατηρήσεων, δηλ., *E[Ot= vk|λ1] = E[Ot= vk|λ2]*, για όλα τα *vk*, απαιτείται

pq + (1 - p) (1 - q) = rs + (1 - r) (1 - s)

ή, με επίλυση ως προς *s*, παίρνουμε

$$s = \frac{p+q-2pq-r}{1-2r}$$

Με την επιλογή (αυθαίρετα) p = 0.6, q = 0.7, r = 0.2, παίρνουμε s = 13/30 = 0.433. Κατά συνέπεια, ακόμα και όταν τα δύο μοντέλα, λ1 και λ2, είναι φαινομενικά πολύ διαφορετικά (δηλ., το Al είναι πολύ διαφορετικό από το A2 και το B1 είναι πολύ διαφορετικό από το B2), τα μοντέλα μπορεί να εμφανίσουν στατιστική ισοδυναμία.

Ο Juang [32] πρότεινε ένα πιθανοτικό μέτρο απόστασης για την ποσοτικοποίηση της ανομοιότητας μεταξύ ζευγών HMMs με αυθαίρετες πυκνότητες παρατήρησης. Επιπλέον, αυτό το μέτρο απόστασης είναι βασισμένο στον αριθμό Kullback-Leibner και ακολουθεί την έννοια της διαγώνιας εντροπίας, απόκλισης ή της σχετικής εντροπίας.

Έστω ότι $\lambda = (A, B, \pi, W, R, U)$ είναι μια κατηγορία HMMs με τις ακόλουθες ιδιότητες: (i) τα αντίστοιχα σύμβολα παρατήρησης που καθορίζονται ($V = \{v1, v2, ..., vM\}$) είναι τα ίδια για όλα τα HMMs μέσα στην κλάση (ii) δύο μεμονωμένα HMMs $\lambda_\varepsilon, \lambda_e \in \lambda$ γενικά, έχουν διαφορετικές πληθικότητες των αντίστοιχων χώρων καταστάσεων ($N(\varepsilon)$ *και* $N(e)$) και έχουν διαφορετικές στοχαστικές μήτρες A και B, διαφορετικές αρχικές κατανομές πιθανότητας π και διαφορετικές μήτρες W, R και U (η διαφορά όσον αφορά μια μήτρα αναφέρεται στις μη ταυτόσημες διαστάσεις ή/και τις μη ταυτόσημες τιμές των αντίστοιχων στοιχείων) (iii) εφόσον η αντίστοιχη αλυσίδα Markov ορίζεται επί ενός πεπερασμένου χώρου καταστάσεων και έχει στάσιμες πιθανότητες μετάβασης που είναι αυστηρά θετικές

$$a_{ij} = e^{w_{ij}} / \sum_{l=1}^{N} e^{w_{il}} > 0, \quad 1 \le i \le N, 1 \le j \le N \tag{3.41}$$

μπορεί να αποδειχθεί ότι η αλυσίδα Markov είναι εργοδική και επομένως έχει μια μοναδική αμετάβλητη κατανομή.

Έστω μια ακολουθία παρατηρήσεων O (μήκους T) που παράγεται από το $\lambda_\varepsilon \in \lambda$. Ορίζουμε μια συνάρτηση $H_T(O, \lambda_e)$ στο λ ως

$$H_T(O, \lambda_e) = \frac{1}{T} \log P(O_T, \lambda_e) \ . \tag{3.42}$$

Σύμφωνα με το θεώρημα ορίου του Petrie για κάθε HMM στο λ, το όριο

$$\lim_{T \to \infty} H_T(O, \lambda_e) = \lim_{T \to \infty} \frac{1}{T} \log P(O_T, \lambda_e) = H(\lambda_\varepsilon, \lambda_e) \tag{3.43}$$

υπάρχει σχεδόν παντού (αφού το λ αποτελείται από εργοδικές αλυσίδες Markov με μοναδικές ομοιόμορφες κατανομές).

Με βάση τα παραπάνω, ορίζουμε ένα πιθανοτικό μέτρο $PD(\lambda_\varepsilon, \lambda_e)$ μεταξύ δύο HMMs λ_e και λ_ε ως

$$PD(\lambda_\varepsilon, \lambda_e) = H(\lambda_\varepsilon, \lambda_\varepsilon) - H(\lambda_\varepsilon, \lambda_e) = \lim_{T \to \infty} \frac{1}{T} \left[\log P(O_T^{(\varepsilon)}, \lambda_\varepsilon) - \log P(O_T^{(\varepsilon)}, \lambda_e) \right] \tag{3.44}$$

το προαναφερθε'ν θεώρημα ορίου εγγυάται την ύπαρξη ενός τέτοιου πιθανοτικού μέτρου και εξασφαλίζει ότι το $PD(\lambda_\varepsilon, \lambda_e)$ είναι μη αρνητικό. Βασικά το (3.44) είναι ένα μέτρο του πόσο καλά το μοντέλο λ_e περιγράφει τις παρατηρήσεις που παράγονται από μοντέλο λ_ε , σε σχέση με το πόσο καλά το μοντέλο λ_ε περιγράφει τις παρατηρήσεις που παράγονται από το ίδιο.

Το μέτρο (3.44) είναι σαφώς μη συμμετρικό. Μια φυσική επέκταση αυτού του μέτρου είναι η ακόλουθη

$$PDS(\lambda_\varepsilon, \lambda_e) = \frac{PD(\lambda_\varepsilon, \lambda_e) + PD(\lambda_e, \lambda_\varepsilon)}{2} \tag{3.45}$$

που είναι η μέση τιμή των δυο μη συμμετρικών πιθανοτικών μέτρων.

Το $PDS(\lambda_e, \lambda_\varepsilon)$ ικανοποιεί τις παρακάτω συνθήκες:

Συμμετρία: $PDS(\lambda_\varepsilon, \lambda_e) = PDS(\lambda_e, \lambda_\varepsilon)$ **(3.46)**

Θετικότητα: $\boldsymbol{PDS\ (\lambda_\varepsilon, \lambda_e) \leq 0}$ **(3.47)**

οπότε αποτελεί ένα πιθανοτικό μέτρο απόστασης/ανομοιότητας.

Η αποτίμηση της (3.45) είναι σχετικά απλή. Μια τυποποιημένη διαδικασία Monte Carlo που βασίζεται σε μια καλή τυχαία γεννήτρια αριθμών, χρησιμοποιείται για να παράγει τις απαιτούμενες ακολουθίες παρατήρησης $O_T^{(\varepsilon)}$ και $O_T^{(e)}$ σύμφωνα με τα HMMs λ_ε και λ_e, αντίστοιχα. Κατόπιν, υπολογίζονται οι πιθανότητες της παρατήρησης των ακολουθιών που έχουν παραχθεί από τα λ_ε και λ_e.

Κεφάλαιο 4
Το υβριδικό νευρωνικό δίκτυο SOHMMM

4.1 Εισαγωγή

Σε αυτό το κεφάλαιο παρουσιάζεται στο σύνολό του ο Self-Organizing Hidden Markov Model Map (SOHMMM) [1] τόσο από μια θεωρητική όσο και από μια αλγοριθμική σκοπιά. Το SOHMMM εισάγει μια σύζευξη μεταξύ των θεωριών και των αλγοριθμικών υλοποιήσεων του Self-Organizing Map (SOM) και του Hidden Markov Model (HMM), δημιουργώντας έτσι ένα υβρίδιο των δύο. Οι αντίστοιχες αρχιτεκτονικές και μεθοδολογίες εκμάθησης συνδυάζονται ώστε να αντιμετωπιστούν οι ολοένα και αυξανόμενες απαιτήσεις των DNA, RNA καθώς και των πρωτεϊνικών μοριακών αλυσίδων. Η εξέταση πολλών από τα πιο ενδιαφέροντα προβλήματα ανάλυσης ακολουθιών της βιολογίας επιτυγχάνεται μέσω του αυτόματου μη-επιβλεπόμενου μηχανισμού εκμάθησης του SOHMMM από ακολουθιακά δεδομένα. Η συνένωση και η σύμπραξη της μη επιβλεπόμενης εκπαίδευσης του SOM και των αλγορίθμων δυναμικού προγραμματισμού του HMM παράγουν έναν νέο on-line αλγόριθμο εκμάθησης καθόδου κλίσης, ο οποίος είναι πλήρως ενσωματωμένος στο SOHMMM. Δεδομένου ότι το SOHMMM πραγματοποιεί μια πιθανοτική ανάλυση ακολουθιών με ελάχιστη ή καμία προγενέστερη γνώση, μπορεί να έχει ποικίλες εφαρμογές στη εύρεση συστάδων, τη μείωση της διάστασης και την απεικόνιση των μεγάλης κλίμακας χώρων ακολουθιών και επίσης, στη σύγκριση, αναζήτηση και ταξινόμηση ακολουθιών.

4.2 Bayesian Inference, εκτίμηση παραμέτρων, και επιλογή μοντέλου

Σε αυτή την ενότητα παρουσιάζουμε την παραγωγή ενός παραμετροποιημένου Hidden Markov Model $\lambda = (A, B, \pi)$ από ένα σώμα δεδομένων O (ακολουθία παρατήρησης) [2]. Από το θεώρημα Bayes έχουμε αμέσως ότι

$$P(\lambda \mid O) = \frac{P(O \mid \lambda) P(\lambda)}{P(O)} \tag{4.1}$$

Η ανωτέρω P(λ) αντιπροσωπεύει την εκτίμησή μας της πιθανότητας το HMM λ να είναι σωστό προτού να λάβουμε οποιαδήποτε δεδομένα. Η μεταγενέστερη $P(\lambda|O)$ αντιπροσωπεύει την ενημερωμένη πίστη μας στην πιθανότητα ότι το HMM λ είναι σωστό λαμβάνοντας υπόψη ότι έχουμε παρατηρήσει το σύνολο δεδομένων O. Ο όρος $P(O|\lambda)$ αναφέρεται ως πιθανότητα (likelihood). Για τεχνικούς λόγους, οι πιθανότητες μπορούν να είναι πολύ μικρές οπότε είναι συχνά ευκολότερο να εργαστούμε με τους αντίστοιχους λογαρίθμους έτσι ώστε

$$\log P(\lambda \mid O) = \log P(O \mid \lambda) + \log P(\lambda) - \log P(O) \tag{4.2}$$

Για να εφαρμόσουμε την (4.2) σε οποιαδήποτε κατηγορία HMMs, θα πρέπει να προσδιορίσουμε την προγενέστερη $P(\lambda)$ και την πιθανότητα δεδομένων P(O|λ). Μόλις καθοριστούν οι όροι αυτοί, η αρχική προσπάθεια μοντελοποίησης έχει ολοκληρωθεί και το μόνο που μένει, αφορά τη θεωρία των πιθανοτήτων.

Ο πιο συνηθισμένος στόχος είναι να βρεθεί ή να προσεγγιστεί το καλύτερο HMM μέσα σε μια κλάση - δηλαδή, να βρεθεί το σύνολο των παραμέτρων που μεγιστοποιούν τη μεταγενέστερη P(λ | O), ή logP(λ | O). Η διαδικασία αυτή καλείται maximum a posteriori

estimation (MAP). Αυτό είναι επίσης ισοδύναμο με την ελαχιστοποίηση - logP (λ | O):

$$- logP(\lambda \mid O) = - logP(O \mid \lambda) - logP(\lambda) + logP(O) \quad . \qquad (4.3)$$

Από μια σκοπιά βελτιστοποίησης, ο λογάριθμος των προηγουμένων παίζει τον ρόλο μιας κανονικοποίησης, δηλαδή, ενός πρόσθετου όρου ποινής που μπορεί να χρησιμοποιηθεί για να επιβάλλει πρόσθετους περιορισμούς, όπως η ομαλότητα. Σημειώνεται πως ο όρος P(O) της (4.3) παίζει τον ρόλο μιας σταθεράς κανονικοποίησης που δεν εξαρτάται από τις παραμέτρους του HMM, και είναι επομένως άσχετος με τη βελτιστοποίηση. Εάν το προγενέστερο P(λ) είναι ομοιόμορφο επί όλων των εξεταζόμενων HMMs, κατόπιν το πρόβλημα περιορίζεται στην εύρεση του μεγίστου του P(O | λ), ή logP(O | λ). Το τελευταίο αποτελεί την εκτίμηση μέγιστης πιθανότητας (ML).

Gradient Descent

Στα περισσότερα ενδιαφέροντα μοντέλα, η συνάρτηση που βελτιστοποιείται είναι σύνθετη και οι μορφές της δεν μπορούν να λυθούν αναλυτικά. Κατά συνέπεια, πρέπει κανείς να προσφύγει σε επαναληπτικές και ενδεχομένως στοχαστικές μεθόδους όπως η κάθοδος κλίσης (gradient descent) ή η simulated annealing [3][4], θεωρώντας επίσης και κατά προσέγγιση ή υπό-βέλτιστες λύσεις. Τέτοιες διαδικασίες έχουν αποδειχθεί να παράγουν λύσεις συγκρίσιμες με εκείνες των τυποποιημένων διαδικασιών reestimation.

Όπως έχει ήδη αναφερθεί, ενδιαφερόμαστε για την εκτίμηση των παραμέτρων, δηλ., την εύρεση του καλύτερου δυνατού HMM λ = λ (X) που ελαχιστοποιεί την μεταγενέστερη $f(x) = -\log P(\lambda(x) \mid O)$, την πιθανότητα $-\log P(O \mid \lambda(x))$. Στην περίπτωση που μια συνάρτηση f είναι διαφορίσιμη, μπορεί κανείς να προσπαθήσει να βρει τα ελάχιστα της με χρήση ενός από τους παλαιότερους αλγορίθμους βελτιστοποίησης, την κάθοδο κλίσης. Όπως δείχνει και το όνομα, η κάθοδος κλίσης είναι μια επαναληπτική διαδικασία, όπου η παράμετρος ενδιαφέροντος, έστω x, ρυθμίζεται σύμφωνα με τον κανόνα

$$x^{next} = x^{now} - \eta \, \nabla f(x)\big|_{x=x^{now}} \qquad (4.4)$$

όπου η είναι το μέγεθος του βήματος, ή ρυθμός εκμάθησης, που μπορεί να καθοριστεί ή ρυθμιστεί κατά τη διάρκεια της διαδικασίας εκμάθησης.

Ενώ η γενική αρχή καθόδου κλίσης είναι απλή, σε σύνθετα παραμετροποιημένα μοντέλα μπορεί να αποτελέσει αφορμή για διαφορετικές υλοποιήσεις, ανάλογα με το πώς υπολογίζεται πραγματικά η κλίση. Στα γραφικά μοντέλα, αυτό απαιτεί συχνά τη διάδοση πληροφορίας «προς τα πίσω». Αυτό συμβαίνει για τον μπροστινός-οπίσθιο (forward-backward) αλγόριθμο των HMM. Προφανώς η έκβαση μιας διαδικασίας καθόδου κλίσης εξαρτάται από την αρχική εκτίμηση. Επιπλέον, εάν η συνάρτηση που βελτιστοποιείται έχει σύνθετο σύνολο τιμών, η κάθοδος κλίσης γενικά θα καταλήξει σε ένα τοπικό ελάχιστο πιθανότερα από ότι σε ένα ολικό. Όποτε είναι εφικτό, επομένως, είναι καλύτερο να εκτελεστεί η διαδικασία αρκετές φορές, με διαφορετικές αρχικές παραμέτρους και ρυθμούς εκμάθησης.

Online Gradient Descent αλγόριθμος για HMMs

Οι εξισώσεις καθόδου κλίσης στην αρνητική λογαριθμική πιθανότητα μπορούν να παραχθούν άμεσα με την εκμετάλλευση μιας χρήσιμης επανα-παραμετροποίησης. Το HMM χρησιμοποιώντας τους κανονικοποιημένους εκθέτες, έχει τη μορφή

$$a_{ij} = \frac{e^{w_{ij}}}{\sum_{l=1}^{N} e^{w_{il}}} \qquad b_j(t) = \frac{e^{r_{jt}}}{\sum_{l=1}^{M} e^{r_{jl}}} \qquad \pi_j = \frac{e^{u_j}}{\sum_{l=1}^{N} e^{u_l}} \tag{4.5}$$

με τα *wij, rjt, uj* ως νέες μεταβλητές [5]. Αυτή η επανα-παραμετροποίηση έχει δύο κύρια πλεονεκτήματα: κατ' αρχάς, η τροποποίηση των νέων μεταβλητών διατηρεί αυτόματα τους τυπικούς στοχαστικούς περιορισμούς στις κατανομές πιθανότητας αρχικής κατάστασης, μετάβασης και εκπομπής και δεύτερον οι πιθανότητες αρχικής κατάστασης, μετάβασης και εκπομπής δεν μπορούν ποτέ να αποκτήσουν τιμή 0. Λόγω αυτών των δύο πλεονεκτημάτων η ανωτέρω επανα-παραμετροποίηση υπερνικά τα εμπόδια του αλγορίθμου Baum-Welch.

Πιο συγκεκριμένα, στην περίπτωση μιας μεμονωμένης ακολουθίας εκπαίδευσης, σε κάθε επανάληψη ο αλγόριθμος Baum-Welch επανα-ρυθμίζει μια παράμετρο μετάβασης ή εκπομπής στην αναμενόμενη συχνότητά της, λαμβάνοντας υπόψη το τρέχον μοντέλο και τα δεδομένα εκπαίδευσης. Είναι σαφές ότι ο αλγόριθμος Baum-Welch μπορεί να οδηγήσει σε απότομα άλματα στον χώρο παραμέτρων και επομένως η διαδικασία δεν είναι κατάλληλη για on-line μάθηση. Ένα άλλο πρόβλημα με τον αλγόριθμο Baum-Welch είναι ότι πιθανότητες με τιμή 0 είναι "απορροφητικές": μόλις τεθεί μια πιθανότητα μετάβασης ή εκπομπής στην τιμή 0, δεν χρησιμοποιείται ποτέ πάλι και παραμένει έκτοτε ίση με 0. Αυτό είναι φυσικά ανεπιθύμητο και συνήθως αποτρέπεται επιβάλλοντας τεχνητά ότι καμία παράμετρος δεν θα είναι μικρότερη από ένα σταθερό μικρό κατώτατο όριο.

Είναι προφανές ότι οι νέες μεταβλητές wij, rjt, uj μπορούν επίσης να οργανωθούν σε μία σειρά από πίνακες, δηλαδή $W = \{w_{ij}\}$, $R = \{r_{jt}\}$ and $U = \{u_j\}$.. Αντίστοιχα, μια πλήρης προδιαγραφή ενός HMM απαιτεί την προδιαγραφή των πληθικοτήτων των δύο χώρων καταστάσεων (συγκεκριμένα N και M), την προδιαγραφή των συμβόλων παρατήρησης, την προδιαγραφή των πιθανοτικών μητρών A και B, την προδιαγραφή της αρχικής κατανομής πιθανότητας π και την προδιαγραφή των μητρών W, R και U. Εφεξής, μπορούμε να χρησιμοποιούμε τον συνοπτικό συμβολισμό για το μοντέλο

$$\lambda = (A, B, \pi, W, R, U). \tag{4.6}$$

Με χρήση των (4.4)-(4.6) μπορεί να εξαχθεί ένας αλγόριθμος για online gradient descent αρνητικής λογαριθμικής πιθανότητας (negative log-likelihood), με w_{ij}, r_{jt}, u_j ως τις παραμέτρους που πρέπει να υπολογιστούν

$$w_{ij}^{(next)} - w_{ij}^{(now)} = -\eta \frac{\partial\left[-\log P(O|\lambda)\right]}{\partial w_{ij}}\Big|_{w_{ij}=w_{ij}^{(now)}} =$$

$$\eta \frac{1}{P(O|\lambda)} \frac{\partial P(O|\lambda)}{\partial w_{ij}}\Big|_{w_{ij}=w_{ij}^{(now)}}, \qquad 1 \le i \le N,\ 1 \le j \le N \tag{4.7}$$

$$r_{jt}^{(next)} - r_{jt}^{(now)} = -\eta \frac{\partial\left[-\log P(O|\lambda)\right]}{\partial r_{jt}}\Big|_{r_{jt}=r_{jt}^{(now)}}$$

$$= \eta \frac{1}{P(O|\lambda)} \frac{\partial P(O|\lambda)}{\partial r_{jt}}\Big|_{r_{jt}=r_{jt}^{(now)}}, \qquad 1 \le j \le N,\ 1 \le t \le M \tag{4.8}$$

$$u_j^{(next)} - u_j^{(now)} = -\eta \frac{\partial\left[-\log P(O|\lambda)\right]}{\partial u_j}\Big|_{u_j=u_j^{(now)}}$$

$$= \eta \frac{1}{P(O|\lambda)} \frac{\partial P(O|\lambda)}{\partial u_j}\Big|_{u_j=u_j^{(now)}}, \qquad 1 \le j \le N \tag{4.9}$$

Με χρήση της (4.7) και του λήμματος 2 (βλ. Παράρτημα Α) η εξίσωση online gradient descent της αρνητικής λογαριθμικής πιθανότητας σε σχέση με το w_{ij} είναι

$$w_{ij}^{(next)} =$$

$$w_{ij} + \eta \frac{a_{ij}}{P(O|\lambda)} \sum_{l=1}^{T-1} \left[\alpha_l(i) b_j(o_{l+1}) \beta_{l+1}(j) - \alpha_l(i) \beta_l(i)\right], \qquad 1 \le i,j \le N \tag{4.10}$$

η δεξιά πλευρά της (4.10) αποτιμάται με βάση τις τρέχουσες τιμές των παραμέτρων των HMMs (A, B, π και W).

Ομοίως, με χρήση της (4.8) και του λήμματος 4 ((βλ. Παράρτημα Α) η online gradient descent εξίσωση της αρνητικής λογαριθμικής πιθανότητας σε σχέση με το r_{jt} είναι

$$r_{jt}^{(next)} = r_{jt} + \eta \frac{1}{P(O|\lambda)} \sum_{l=1}^{T} \left[I\left\{o_l = t|\lambda\right\} \alpha_l(j) \beta_l(j) - b_j(t) \alpha_l(j) \beta_l(j)\right],$$

$$1 \le j \le N,\ 1 \le t \le M \tag{4.11}$$

η δεξιά πλευρά της (4.11) αποτιμάται επίσης με βάση τις τρέχουσες τιμές των παραμέτρων των HMMs (A, B, π και R).

Τελικά, με χρήση της (4.9) και του λήμματος 6 (βλ. Παράρτημα Α) η online gradient descent εξίσωση της αρνητικής λογαριθμικής πιθανότητας σε σχέση με το u_j είναι

$$u_j^{(next)} = u_j + \eta \frac{\pi_j}{P(O|\lambda)} \left[b_j(o_1) \beta_1(j) - P(O|\lambda)\right]$$

$$= u_j + \eta \pi_j \left[\frac{b_j(o_1) \beta_1(j)}{P(O|\lambda)} - 1\right], \ 1 \le j \le N \tag{4.12}$$

η δεξιά πλευρά της (4.12) αποτιμάται επίσης με βάση τις τρέχουσες τιμές των παραμέτρων των HMMs (A, B, π και U).

Σημειώνεται ότι η συμβολή κάθε ακολουθίας πρέπει να σταθμιστεί από το αντίστροφο της πιθανότητας της. Αυτό μπορεί να φανεί ότι κάπως παράδοξο καταρχάς, αλλά γίνεται κατανοητό εύκολα. Εάν το HMM αντιστοιχίσει μια μικρή πιθανότητα σε μια δεδομένη

παρατηρηθείσα ακολουθία, αυτό σημαίνει ότι το HMM δεν συμπεριφέρεται κατάλληλα σε εκείνη την ακολουθία και επομένως το αντίστοιχο σήμα λάθους πρέπει να είναι υψηλό.

Γενικά, ο αλγόριθμος που εισήχθη ανωτέρω πρέπει να είναι χρήσιμος στις περιπτώσεις όπου η ομαλότητα ή/και η on-line εκμάθηση είναι σημαντικές. Αυτό θα μπορούσε να συμπεριλάβει: (i) μεγάλα μοντέλα με πολλές παραμέτρους και σχετικά λιγοστά δεδομένα που μπορούν να είναι πιο επιρρεπή σε υπερειδίκευση (overfitting), τοπική παγίδευση ελάχιστων και άλλες παθολογικές συμπεριφορές όταν εκπαιδεύονται με τους ασυνεχείς ή batch κανόνες (ii) αναλογικές φυσικές εφαρμογές όπου μόνο συνεχείς κανόνες εκμάθησης μπορούν να χρησιμοποιηθούν (iii) και όλες οι καταστάσεις όπου η αποθήκευση των παραδειγμάτων για batch εκμάθηση δεν είναι επιθυμητή.

4.3 Self-Organizing Hidden Markov Model Map (SOHMMM)

Στόχος της διερευνητικής ανάλυσης δεδομένων είναι να παραχθούν απλουστευμένες και συνοπτικές περιγραφές μεγάλων συνόλων δεδομένων. Η οργάνωση των δεδομένων σε συστάδες είναι μια τυπική μέθοδος, ενώ μια άλλη εναλλακτική λύση είναι να προβληθούν τα υψηλής διάστασης σύνολα δεδομένων ως σημεία σε έναν χαμηλής διάστασης, συνήθως δισδιάστατο χώρο.

Ο σκοπός των μεθόδων προβολής είναι να μειωθεί η διάσταση των διανυσμάτων δεδομένων. Οι προβολές αντιπροσωπεύουν τα στοιχεία των δεδομένων εισόδου σε ένα χαμηλής διάστασης χώρο κατά τέτοιο τρόπο ώστε οι συστάδες και οι μετρικές σχέσεις των δεδομένων να διατηρείται όσο το δυνατόν πιο πιστά. Οι προβολές μπορούν επίσης να χρησιμοποιηθούν για να απεικονίσουν τα σύνολα δεδομένων εάν επιλέγεται μια αρκετά μικρή διάσταση (το πολύ μέχρι τρεις διαστάσεις) για τον χώρο εξόδου.

Ο αυτο-οργανούμενος χάρτης συνδυάζει τις διαδικασίες οργάνωσης των δεδομένων σε συστάδες και προβολής. Το SOM αντιπροσωπεύει τα δεδομένα εισόδου με έναν τακτικό τρόπο σε ένα δισδιάστατο κανονικό πλέγμα, και τα στοιχεία του μοντέλου αντιστοιχίζονται στη συνέχεια με τους κόμβους/νευρώνες του πλέγματος. Οι κύριες εφαρμογές του SOM είναι έτσι στην απεικόνιση σύνθετων δεδομένων σε δισδιάστατο χώρο εξόδου και τη δημιουργία αφαιρέσεων όπως σε πολλές τεχνικές clustering.

Μια εξέλιξη στην ιδέα του SOM είναι ο Self-Organizing Hidden Markov Model Map, μια νέα προσέγγιση ως προς την απεικόνιση και εύρεση συστάδων ακολουθιών παρατηρήσεων. Στη βασική του μορφή παράγει έναν γράφο της κατανομής απόστασης/πιθανότητας των δεδομένων εισόδου. Μετατρέπει τις μη γραμμικές στατιστικές σχέσεις μεταξύ των δεδομένων ακολουθιών παρατηρήσεων σε απλές γεωμετρικές σχέσεις των απεικονίσεων τους σε ένα χαμηλής-διάστασης χώρο, συνήθως ένα κανονικό δισδιάστατο πλέγμα κόμβων. Δεδομένου ότι το SOHMMM με αυτόν τον τρόπο συμπιέζει πληροφορία συντηρώντας τις σημαντικότερες τοπολογικές και στατιστικές σχέσεις των αρχικών δεδομένων στο πλέγμα, μπορεί επίσης να θεωρηθεί ότι παράγει κάποιο είδος αφαίρεσης. Αυτές οι δύο πτυχές, η απεικόνιση και η αφαίρεση σε συνδυασμό με με τη δημιουργία συστάδων, μπορούν να χρησιμοποιηθούν με διάφορους τρόπους σε σύνθετες εργασίες.

Οι προσομοιώσεις που πραγματοποιήθηκαν κατά τη διάρκεια πολλών ετών από έναν μεγάλο αριθμό επιστημόνων έχουν δείξει ότι τα καλύτερα αποτελέσματα αυτο-οργάνωσης λαμβάνονται εάν οι ακόλουθες δύο διαδικασίες εφαρμόζονται στις καθαρότερες μορφές τους [6]: (1) αποκωδικοποίηση εκείνου του νευρώνα («νικητή») που ταιριάζει καλύτερα με το δοσμένο δείγμα των δεδομένων εισόδου. (2) προσαρμοζόμενη βελτίωση του ταιριάσματος στη γειτονιά των νευρώνων γύρω από τον «νικητή».

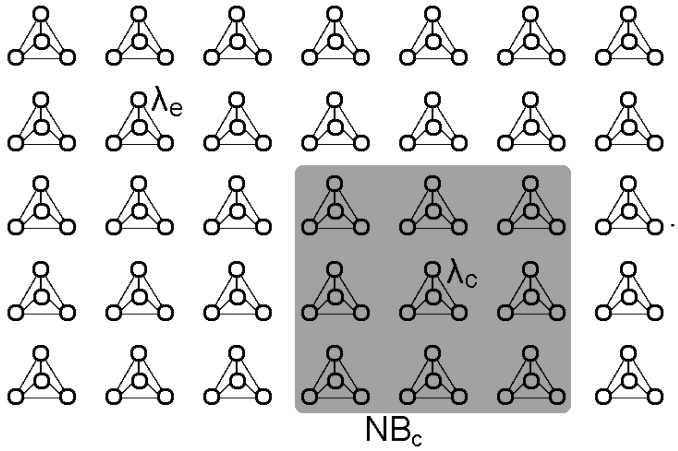
Το SOHMMM μπορεί να περιγραφεί τυπικά ως μια μη γραμμική, διατεταγμένη, ομαλή χαρτογράφηση των δεδομένων ακολουθιών παρατηρήσεων επάνω στα στοιχεία ενός κανονικού, χαμηλής διάστασης πίνακα Η χαρτογράφηση εφαρμόζεται με τον ακόλουθο τρόπο, ο οποίος μοιάζει με τις δύο προαναφερθείσες διαδικασίες. Υποθέτουμε αρχικά ότι Ο είναι μια ακολουθία παρατήρησης. Με κάθε στοιχείο στον πίνακα SOHMMM αντιστοιχίζουμε ένα HMM λ_e. Υποθέτοντας την πιθανότητα της Ο, δοθέντος του λ_e (likelihood), συμβολιζόμενη ως $P(O \mid \lambda_e)$, η εικόνα μιας ακολουθίας παρατήρησης εισόδου Ο στον πίνακα SOHMMM ορίζεται ως το στοιχείο του πίνακα λ_e που ταιριάζει καλύτερα με την Ο, δηλαδή που έχει τον δείκτη

$$c = \arg\max_e \{P(O \mid \lambda_e)\}$$

ή ισοδύναμα

$$c = \arg\min_e \{-\log P(O \mid \lambda_e)\}$$

αφού ο αρνητικός λογάριθμος είναι μια μονότονα φθίνουσα συνάρτηση



Εικόνα 4.1. Παράδειγμα SOHMMM πλέγματος

Ο στόχος μας είναι να καθορίσουμε το λ_e με τέτοιο τρόπο ώστε η χαρτογράφηση να είναι διατεταγμένη και περιγραφική της κατανομής των Ο. Θεωρούμε την εικόνα 4.1 όπου απεικονίζεται ένας δισδιάστατος διατεταγμένος πίνακας κόμβων, καθένας από τους οποίους έχει ένα HMM λ_e αντιστοιχισμένο με αυτόν. Επιπλέον, θεωρούμε το σύνολο γειτονιάς NB_c γύρω από το μοντέλο λ_c, το οποίο ταιριάζει καλύτερα με την Ο. Εδώ, το NB_c αποτελείται από όλους τους νευρώνες μέχρι μια ορισμένη ακτίνα στο πλέγμα από το νευρώνα c. Ο επόμενος στόχος είναι η ρύθμιση των παραμέτρων, όλων των HMMs μέσα στο NB_c, ώστε να ελαχιστοποιηθεί η $-\log P(O|\lambda_e)$. Στην πραγματικότητα, προσπαθούμε να βελτιστοποιήσουμε τις παραμέτρους κάθε λ_e στο NB_c ώστε να περιγραφεί καλύτερα πώς μια δεδομένη παρατήρηση εμφανίζεται. Αυτό επιτυγχάνεται με την χρήση της καθόδου κλίσης, τον ομαλό, online αλγόριθμο εκμάθησης που περιγράφηκε προηγουμένως.

4.3.1 Το βασικό SOHMMM

Έστω $O = o_1o_2..o_T$ μια ακολουθία παρατήρησης όπου κάθε παρατήρηση o_t λαμβάνει μια τιμή από το αλφάβητο $F = \{f_1, f_2, \ldots, f_G\}$, και T είναι ο αριθμός των παρατηρήσεων στην ακολουθία. Επιπρόσθετα, έστω $\lambda = (A, B, \pi, W, R, U)$ μια κλάση από HMMs έτσι ώστε τα αντίστοιχα σύμβολα παρατήρησης που καθορίζονται ($V = \{v_1, v_2, \ldots, v_M\}$) να είναι υπερσύνολο του αλφάβητου F. Μια περαιτέρω υπόθεση που γίνεται είναι ότι δύο διακριτά HMMs λ_ε, λ_e, γενικά, έχουν διαφορετικές πληθικότητες των αντίστοιχων χώρων καταστάσεων ($N^{(\varepsilon)} \neq N^{(e)}$ και $M^{(\varepsilon)} \neq M^{(e)}$), διαφορετικά σύμβολα παρατηρήσεων ($V^{(\varepsilon)} \neq V^{(e)}$), διαφορετικές στοχαστικές μήτρες A και B, διαφορετικές αρχικές κατανομές πιθανότητας π και διαφορετικές μήτρες W, R and U (η διαφορά όσον αφορά μια μήτρα αναφέρεται στις μη ταυτόσημες διαστάσεις ή/και τις μη ταυτόσημες τιμές των αντίστοιχων στοιχείων).

Έστω η εικόνα 4.2. Το SOHMMM καθορίζει εδώ μια χαρτογράφηση από το διάστημα ακολουθιών παρατηρήσεων εισόδου επάνω σε έναν δισδιάστατο πίνακα νευρώνων.
Με κάθε νευρώνα e αντιστοιχίζεται ένα HMM λe, αποκαλούμενο επίσης HMM αναφοράς. Ο τύπος πλέγματος του πίνακα μπορεί να καθοριστεί ορθογώνιος, εξαγωνικός ή ακόμα και μη κανονικός. Στην απλούστερη περίπτωση, μια ακολουθία O συνδέεται με όλα τα HMMs παράλληλα. Σε ένα αφηρημένο σχήμα, μπορεί κανείς να θεωρήσει ότι η είσοδος O, με τη βοήθεια κάποιων μηχανισμών παράλληλου υπολογισμού, συγκρίνεται με όλα τα λe, και η θέση της καλύτερης αντιστοιχίας όσον αφορά την αρνητική λογαριθμική πιθανότητα, ορίζεται ως θέση της απόκρισης. Στην πραγματικότητα, το ακριβές μέγεθος της απόκρισης δεν χρειάζεται να καθοριστεί: η είσοδος χαρτογραφείται απλά επάνω σε αυτήν την θέση, όπως σε ένα σύνολο αποκωδικοποιητών. Μπορούμε στη συνέχεια να υποστηρίξουμε ότι το SOHMMM είναι μια μη γραμμική προβολή των ακολουθιών παρατηρήσεων εισόδου επάνω στο δισδιάστατο πλέγμα. Κατά συνέπεια, ο ορισμός του καλύτερα ταιριάζοντος HMM, είναι ο ακόλουθος

$$c = \arg\min_{e} \left\{ -\log P(O|\lambda_e) \right\} \tag{4.13}$$

Κατά τη διάρκεια της εκμάθησης, ή της διαδικασίας κατά την οποία διαμορφώνεται η μη γραμμική προβολή, εκείνα τα HMMs που είναι τοπολογικά κοντά στον πίνακα μέχρι μια ορισμένη γεωμετρική απόσταση, θα ενεργοποιήσουν το ένα το άλλο ώστε να "μάθουν" κάτι από την ίδια είσοδο O. Αυτό θα οδηγήσει σε μια τοπική χαλάρωση ή ένα φαινόμενο smoothing στις παραμέτρους των HMMs σε αυτήν την γειτονιά, που στη συνεχή εκμάθηση οδηγεί σε μια ολική διάταξη. Προκειμένου να ρυθμιστούν οι παράμετροι των HMMs για να μεγιστοποιηθεί η αντίστοιχη αρνητική λογαριθμική πιθανότητα ($-\log P(O|\lambda_e)$) ακολουθούμε μια υβριδική προσέγγιση με τη συγχώνευση του αρχικού αλγορίθμου SOM και του αλγορίθμου καθόδου κλίσης για την αρνητική λογαριθμική πιθανότητα. Ο προκύπτων αλγόριθμος εκμάθησης SOHMMM είναι μια επαναληπτική διαδικασία, όπου οι παράμετροι ενδιαφέροντος, δηλαδή $W^{(e)}$, $R^{(e)}$ και $U^{(e)}$, ρυθμίζονται σύμφωνα με τους κανόνες:

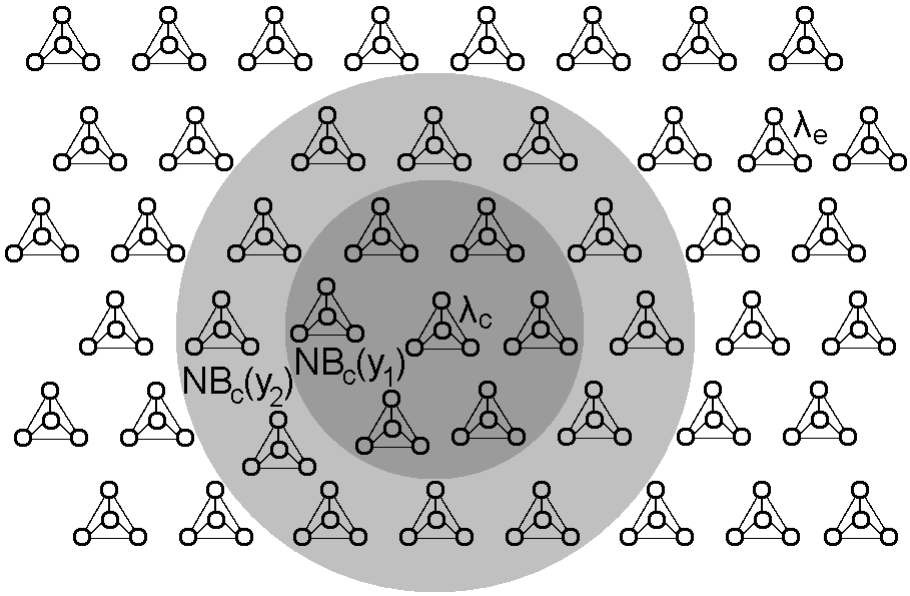
$$w_{ij}^{(e)}(y+1) = w_{ij}^{(e)}(y) + h_{ce}(y)\left[P(O|\lambda)^{-1} a_{ij} \sum_{l=1}^{T-1} \left[\alpha_l(i) b_j(o_{l+1}) \beta_{l+1}(j) - \alpha_l(i)\beta_l(i) \right]\Big|_{\lambda_e, y} \right] \tag{4.14}$$

$$r_{jt}^{(e)}(y+1) =$$

$$r_{jt}^{(e)}(y)+h_{ce}(y)\left[P(O|\lambda)^{-1}\sum_{l=1}^{T}\left[I\{o_l=t|\lambda\}\alpha_l(j)\beta_l(j)-b_j(t)\alpha_l(j)\beta_l(j)\right]\Big|_{\lambda_e,y}\right] \quad \textbf{(4.15)}$$

$$u_j^{(e)}(y+1) = u_j^{(e)}(y)+h_{ce}(y)\left[\pi_j\left[P(O\mid\lambda)^{-1}b_j(o_1)\beta_1(j)-1\right]\Big|_{\lambda_e,y}\right] \quad \textbf{(4.16)}$$

όπου $y = 0, 1, 2, \ldots$ είναι μια ακέραια διακριτού χρόνου μεταβλητή.



Εικόνα 4.2. ένα παράδειγμα τοπολογικής γειτονιάς ($y_1 < y_2$)

Στη διαδικασία χαλάρωσης, η συνάρτηση $h_{ce}(y)$ έχει έναν πολύ πρωταρχικό ρόλο: ενεργεί ως η αποκαλούμενη συνάρτηση γειτονιάς. Για ευκολία, είναι απαραίτητο η $h_{ce}(y)\to 0$ όταν $y\to\infty$. Συνήθως $h_{ce}(y) = h(\|\delta_c - \delta_e\|, t)$, όπου δ_c, δ_e είναι τα διανύσματα θέσης των HMMs c και e, αντίστοιχα, στον πίνακα. Με την αύξηση της διαφοράς $\|\delta_c - \delta_e\|$, $h_{ce}\to 0$. Το μέσο πλάτος και η μορφή της $h_{ce}(y)$ καθορίζουν την ελαστικότητα της επιφάνειας που θα προσαρμοστεί στα δεδομένα εισόδου.

Στη βιβλιογραφία, εμφανίζονται συχνά δύο απλές επιλογές για την $h_{ce}(y)$. Η απλούστερη αναφέρεται σε ένα σύνολο σημείων γειτονιάς γύρω από το HMM c (εικόνα 4.2). Έστω NB_c ο συμβολισμός του συνόλου αυτού (παρατηρούμε ότι μπορούμε να καθορίσουμε $NB_c = NB_c(y)$ ως συνάρτηση του χρόνου), με τον οποίο $h_{ce}(y) = \eta(y)$ αν το e ανήκει στο NB_c και $h_{ce}(y) = 0$ αν το e δεν ανήκει στο NB_c. Η τιμή του $\eta(y)$ έχει τον ρόλο ενός ρυθμού εκπαίδευσης ($0 < \eta(y) < 1$). Τόσο το $\eta(y)$ όσο και η ακτίνα της γειτονιάς NB_c μειώνονται συνήθως μονότονα στον χρόνο (τουλάχιστον στην φάση διάταξης).

Μια άλλη ευρέως χρησιμοποιούμενη και πιο ομαλή συνάρτηση γειτονιάς είναι η συνάρτηση Gauss

$$h_{ce}(y) = \eta(y)\exp\left(-\frac{\|\delta_c-\delta_e\|^2}{2\sigma^2(y)}\right) \quad \textbf{(4.14)}$$

όπου η είναι ένας ρυθμός εκμάθησης και η παράμετρος σ καθορίζει το πλάτος του πυρήνα, αντιστοιχώντας στην ακτίνα $NB_c(y)$ ανωτέρω. Τόσο το η όσο και το σ είναι κάποιες μονότονα φθίνουσες συναρτήσεις του χρόνου.

Ο αλγόριθμος εκμάθησης του SOHMMM που περιγράφηκε είναι αντιπροσωπευτικός πολλών εναλλακτικών μορφών και μπορεί προφανώς να δώσει αφορμή για διάφορες παραλλαγές και διαφορετικές υλοποιήσεις. Ο αλγόριθμος εκμάθησης SOHMMM έχει αναλυθεί στη γενικότερη μορφή του, αποφεύγοντας ωστόσο μια ασαφή περιγραφή.

4.3.2 Θέματα υλοποίησης για το SOHMMM

Το SOM έχει παίξει κεντρικό ρόλο στον χώρο των νευρωνικών δικτύων που βασίζονται στην οργάνωση σε συστάδες. Αν και ο αλγόριθμος του SOM αναπτύχθηκε, κατά πρώτο λόγο, για τη δημιουργία συστάδων και την απεικόνιση μη γραμμικών σχέσεων πολυδιάστατων δεδομένων, έχει αποδειχθεί ότι η ανάλυση δεδομένων που παρέχει έχει ένα πολύ βασικό ρόλο στην κατανόηση των διάφορων φαινομένων. Τα SOM έχουν χρησιμοποιηθεί αποτελεσματικά σε ποικίλους τομείς εφαρμογών. Επιπλέον, τα HMMs έχουν πολύ πλούσια μαθηματική δομή και ως εκ τούτου μπορούν να αποτελέσει τη θεωρητική βάση για τη χρήση σε ένα ευρύ φάσμα εφαρμογών. Επίσης, τα μοντέλα, όταν εφαρμόζονται κατάλληλα, λειτουργούν πολύ καλά στην πράξη.

Η συζήτηση στα προηγούμενα δύο τμήματα έχει εξετάσει πρώτιστα τη θεωρία του SOHMMM και διάφορων πιθανών παραλλαγών στη μορφή του μοντέλου. Προφανώς, το SOHMMM είναι ένα υβρίδιο του αλγορίθμου SOM και της θεωρίας των HMMs που έχουν και τα δύο μελετηθεί, αναλυθεί και εφαρμοστεί ευρέως. Συνεπώς, με η γνώση που έχει αποκτηθεί ως τώρα από αυτούς τους δύο ερευνητικούς τομείς, είμαστε σε θέση να εξετάσουμε διάφορα πρακτικά ζητήματα υλοποίησης [2], [6], [7]. Αυτά τα ζητήματα περιλαμβάνουν τις αρχικές εκτιμήσεις των παραμέτρων του HMM, της κανονικοποίησης, της επιλογής του μεγέθους των HMM, τη μορφή του πλέγματος, τη συνάρτηση γειτονιάς, την προδιαγραφή του ρυθμού εκμάθησης, τον αριθμό των βημάτων εκμάθησης και της βελτίωσης των σπάνιων περιπτώσεων. Για μερικά από αυτά τα ζητήματα μπορούμε να ορίσουμε σχεδόν ακριβείς λύσεις, ενώ για άλλα ζητήματα μπορούμε μόνο να παρέχουμε κάποια εμπειρία που έχει αποκτηθεί από τη χρήση των SOM και HMMs κατά τη διάρκεια των τελευταίων ετών.

Αρχικές εκτιμήσεις των παραμέτρων των HMMs

Θεωρητικά, οι εξισώσεις reestimation πρέπει να δώσουν τιμές στις παραμέτρους HMM που αντιστοιχούν σε ένα τοπικό μέγιστο της συνάρτησης πιθανότητας. Μια ερώτηση κλειδί επομένως είναι πώς επιλέγουμε τις αρχικές εκτιμήσεις των παραμέτρων HMM έτσι ώστε το τοπικό μέγιστο να είναι το ολικό μέγιστο της συνάρτησης πιθανότητας.

Βασικά, δεν υπάρχει καμία απλή ή άμεση απάντηση στην ανωτέρω ερώτηση. Αντί αυτού, η εμπειρία έχει δείξει ότι είτε τυχαίες (που να υπακούν τους τυπικούς στοχαστικούς περιορισμούς) είτε ομοιόμορφες αρχικές εκτιμήσεις των παραμέτρων π και A είναι επαρκείς ώστε να παράγουν χρήσιμα reestimates αυτών των παραμέτρων σχεδόν σε όλες τις περιπτώσεις. Εντούτοις, για τις παραμέτρους B, η εμπειρία έχει δείξει ότι καλές αρχικές εκτιμήσεις είναι χρήσιμες στην περίπτωση διακριτών συμβόλων. Υπάρχουν διάφοροι τρόποι με τους οποίους προγενέστερη πληροφορία μπορεί να ενσωματωθεί στη σχεδίαση των παραμέτρων των HMM. Τέτοιες αρχικές εκτιμήσεις μπορούν να ληφθούν με διάφορους

τρόπους, συμπεριλαμβανομένης της μη αυτόματης κατάτμησης των ακολουθιών παρατήρησης σε καταστάσεις με τον υπολογισμό του μέσου όρου των παρατηρήσεων μέσα στις καταστάσεις, της κατάτμησης μέγιστης πιθανότητας των παρατηρήσεων με υπολογισμό του μέσου όρου, και κατάτμηση κ-μέσων με οργάνωση σε συστάδες.

Μια τέτοια προσέγγιση στην περίπτωση του SOHMMM πρέπει να αποφευχθεί για τρεις κύριους λόγους. Κατ' αρχάς, η προ-επεξεργασία όσον αφορά προγενέστερη πληροφορία εισάγει συμπληρωματικό υπολογιστικό κόστος. Δεύτερον, οι χειρωνακτικές ή ακόμα και ευριστικές εκτιμήσεις παρεκκλίνουν αρκετά από την προσέγγιση μη επιβλεπόμενης μάθησης του SOHMMM. Τρίτον, επιλέγουμε τυχαίους αριθμούς για όλες τις παραμέτρους (σύμφωνα με τους τυπικούς στοχαστικούς περιορισμούς) για να καταδείξουμε ότι εκκινώντας από μια αυθαίρετη αρχική κατάσταση, μακροπρόθεσμα τα HMMs αναφοράς θα αποκτήσουν τελικά διατεταγμένες τιμές. Σε μια άλλη διατύπωση, αρχικά μη ταξινομημένα HMMs θα διαταχτούν τελικά, σε μερικές εκατοντάδες βήματα. Αυτή είναι μια βασική επίδραση της διαδικασίας αυτο-οργάνωσης.

Ένα άλλο πειραματικά ελεγμένο γεγονός για το HMM είναι ότι είναι σημαντικό να τεθούν περιορισμοί για μερικές από τις παραμέτρους υπολογισμού προκειμένου να αποτραπεί το να γίνουν πάρα πολύ μικρές. Παραδείγματος χάριν, ο περιορισμός το $b_j(k)$ να είναι μεγαλύτερο ή ίσο κάποια ελάχιστη τιμή, είναι απαραίτητος για να εξασφαλιστεί ότι ακόμα κι αν το κ-οστό σύμβολο δεν εμφανίστηκε ποτέ σε κάποια κατάσταση j στο σύνολο παρατηρήσεων εκπαίδευσης, υπάρχει πάντα μια πεπερασμένη πιθανότητα εμφάνισής του κατά την αποτίμηση ενός άγνωστου συνόλου παρατηρήσεων.

Κλίμακα

Οι πιθανότητες $P(o_1o_2...o_t \mid \lambda)$ είναι συνήθως πολύ μικρές, δεδομένου ότι είναι ίσες με το γινόμενο πολλών πιθανοτήτων μετάβασης και εκπομπής, με καθεμία μικρότερη του 1. Μπορεί να δειχθεί ότι καθώς το t αρχίζει να αυξάνεται (π.χ. 10 ή περισσότερο) κάθε όρος $\alpha_i(t)$ αρχίζει να τείνει εκθετικά στο μηδέν. Για το αρκετά μεγάλο t (π.χ. 100 ή περισσότερο) η δυναμική περιοχή του υπολογισμού των $\alpha_i(t)$ θα υπερβεί την ακρίβεια κάθε υπολογιστικής μηχανής, ακόμη και στη διπλής ακρίβειας. Κατά συνέπεια, ο υπολογισμός των $\alpha_i(t)$ αποδεικνύεται πέρα από την ακρίβεια μηχανών καθώς το t αυξάνεται. Μια παρόμοια παρατήρηση μπορεί να γίνει για τις οπίσθιες μεταβλητές $\beta_i(t)$ καθώς το t μειώνεται. Επομένως, στην εφαρμογή των αλγορίθμων εκμάθησης, και ειδικά στην περίπτωση των μπροστινών και οπίσθιων διαδικασιών, βρισκόμαστε αντιμέτωποι με ζητήματα ακρίβειας. Αυτά μπορούν να εξεταστούν με τη χρήση μιας διαδικασίας δημιουργίας κλίμακας, όπου οι μπροστινές και οπίσθιες μεταβλητές κανονικοποιούνται (από έναν κατάλληλο συντελεστή που εξαρτάται μόνο από το t) κατά τη διάρκεια της διάδοσης προκειμένου να αποφευχθεί η υποχείλιση. Οι κλίμακες των μπροστινών και οπίσθιων μεταβλητών καθορίζονται με έναν συμπληρωματικό τρόπο έτσι ώστε οι εξισώσεις εκμάθησης να παραμένουν ουσιαστικά αμετάβλητες κάτω από την κλίμακα.

Επιλογή του μεγέθους των HMMs

Ένα άλλο σημαντικό ζήτημα στην υλοποίηση των HMMs είναι η επιλογή του μεγέθους του μοντέλου (αριθμός καταστάσεων). Δυστυχώς, δεν υπάρχει κανένας απλός, θεωρητικά σωστός, τρόπος για μια τέτοια επιλογή Αυτή η επιλογή πρέπει να γίνει ανάλογα με τα δεδομένα εισόδου που μοντελοποιούνται. Ειδικά, στην περίπτωση όπου οι καταστάσεις της αλυσίδας Markov στο HMM έχουν κάποια φυσική ή βιολογική ερμηνεία, τεχνικές

αναζήτησης πλέγματος είναι ο μόνος τρόπος για τον προσδιορισμό της αρχιτεκτονικής των HMM.

Συχνά στη βιβλιογραφία, το ζήτημα του αριθμού των καταστάσεων που χρησιμοποιούνται σε κάθε HMM οδηγεί σε μια ορισμένη σχολή σκέψης. Η ιδέα είναι να αφεθεί ο αριθμός καταστάσεων να αντιστοιχεί κατά προσέγγιση στο μέσο αριθμό παρατηρήσεων μεταξύ των ακολουθιών παρατηρήσεων εισόδου. Στην πράξη, αυτή η απλή προσέγγιση φαίνεται να λειτουργεί αρκετά καλά.

Μορφή του πλέγματος

Για οπτική επιθεώρηση, το εξαγωνικό δικτυωτό πλέγμα είναι προτιμητέο, επειδή δεν ευνοεί τις οριζόντιες και κάθετες κατευθύνσεις τόσο όσο ο ορθογώνιος πίνακας. Οι άκρες του πίνακας οφείλουν να είναι ορθογώνιες παρά τετραγωνικές, επειδή το ελαστικό δίκτυο που διαμορφώνεται από τα HMMs λε πρέπει να προσανατολιστεί και να σταθεροποιηθεί κατά μήκος αυτών των δύο διαστάσεων. Παρατηρήστε ότι εάν ο πίνακας ήταν κυκλικός, δεν θα είχε κανένα σταθερό προσανατολισμό στο χώρο των ακολουθιών παρατηρήσεων εισόδου έτσι οποιαδήποτε επιμήκης μορφή είναι προτιμητέα.

Συνάρτηση γειτονιάς

Πρόσθετη προσοχή, εντούτοις, απαιτείται στην επιλογή του μεγέθους της $NB_c = NB_c(y)$. Εάν η γειτονιά είναι πάρα πολύ μικρή αρχικά, ο χάρτης δεν θα διαταχτεί συνολικά. Αντί' αυτού τα διάφορα είδη "μωσαϊκών" θα φαίνονται στον χάρτη, μεταξύ των οποίων οι αλλαγές κατεύθυνσης διάταξης θα είναι ασυνεχείς. Αυτό το φαινόμενο μπορεί να αποφευχθεί με την έναρξη με αρκετά ευρεία γειτονιά $NB_c = Nb_c(0)$, αφήνοντάς τη να συρρικνωθεί με το χρόνο. Η αρχική ακτίνα NBc μπορεί ακόμη και να είναι περισσότερο από τη μισή διάμετρο του δικτύου. Κατά τη διάρκεια των πρώτων 1000 βημάτων περίπου, όταν πραγματοποιείται η κατάλληλη διάταξη, και το η = η(y) είναι αρκετά μεγάλο, η ακτίνα NBc μπορεί να συρρικνωθεί γραμμικά ως, για παράδειγμα, την μονάδα κατά τη διάρκεια της φάσης tuning, η NBc μπορεί να περιέχει τους πλησιέστερους γείτονες του HMM c. Το τελικό συμπέρασμα είναι ότι εφ' όσον αρχίζει κανείς διαδικασία αυτό-οργάνωσης με μια ευρεία συνάρτηση γειτονιάς (δηλ. με μια ευρεία ακτίνα γειτονιάς NBc, ή μια ευρεία τυπική απόκλιση της $h_{ce}(0)$ ίδιου μεγέθους με τη μισή της μεγαλύτερης διάστασης του πίνακα) δεν υπάρχει συνήθως κανένας κίνδυνος να καταλήξουμε σε ασταθείς διαμορφώσεις του χάρτη.

Προσδιορισμός του ρυθμού εκμάθησης

Η χρονική περίοδος κατά τη διάρκεια της οποίας επιτυγχάνεται μια χονδρική διάταξή στο SOM είναι συνήθως σχετικά σύντομη, της τάξεως των 1000 βημάτων, ενώ το μεγαλύτερο μέρος του χρόνου υπολογισμού αφιερώνεται για την τελική φάση σύγκλισης, προκειμένου να επιτευχθεί μια αρκετά καλή στατιστική ακρίβεια. Δεν είναι σαφές πώς ο παράγοντας εκμάθησης θα πρέπει να βελτιστοποιηθεί κατά τη διάρκεια της πρώτης φάσης, αφού το πλάτος της συνάρτησης γειτονιάς αλλάζει με αυτό τον τρόπο επίσης, κάτι που περιπλέκει την κατάσταση.

Αντίθετα, δεδομένου ότι κατά τη διάρκεια της τελικής φάσης σύγκλισης κρατάμε συνήθως το πλάτος της γειτονιάς σταθερό, φαίνεται δυνατό να καθοριστεί κάποιο είδος βέλτιστου νόμου για την ακολουθία {η(y)} κατά τη διάρκεια της φάσης tuning.

Εάν οι αρχικές τιμές έχουν επιλεχτεί τυχαία, για περίπου τα πρώτα 1000 βήματα, το η(y) πρέπει να έχει εύλογα υψηλές τιμές (κοντά στην μονάδα), που μειώνονται έκτοτε μονότονα. Μια ακριβής χρονική συνάρτηση δεν είναι σημαντική η η = η(y) μπορεί να είναι γραμμική, εκθετική, ή αντιστρόφως ανάλογη ως προς το y. Για παράδειγμα, η(y) = 0.9(1-y/1000) μπορεί να είναι μια λογική επιλογή. Η διάταξη εμφανίζεται κατά τη διάρκεια της αρχικής περιόδου, ενώ τα υπόλοιπα βήματα απαιτούνται μόνο για το συντονισμό του χάρτη. Μετά από τη φάση διάταξης η η = η(y) πρέπει να αποκτήσει τιμές (π.χ. της τάξης λιγότερο από 0.01) κατά τη διάρκεια μιας μεγάλης περιόδου. Ούτε σε αυτή την περίπτωση είναι κρίσιμο εάν ο νόμος για το η(y) μειώνεται γραμμικά ή εκθετικά κατά τη διάρκεια της φάσης tuning. Στους πολύ μεγάλους χάρτες, εντούτοις, μπορεί να είναι σημαντικό να ελαχιστοποιηθεί ο συνολικός χρόνος εκμάθησης. Κατόπιν η επιλογή ενός βέλτιστου νόμου για την η(y) μπορεί να είναι κρίσιμη. Συνήθως οι βέλτιστοι νόμοι για την η(y) ανήκουν στην οικογένεια των αντιστρόφως ανάλογων συναρτήσεων του y. Αποτελεσματικές επιλογές για αυτές τις συναρτήσεις και για τις παραμέτρους τους έχουν ως τώρα προσδιοριστεί εμπειρικά. Ένας τέτοιος μ-εμπειρικός νόμος με την $\eta(y)$ κοινή για όλους τους νευρώνες είναι της μορφής

$$\eta(y) = \frac{\Gamma}{y + \Phi} \tag{4.15}$$

όπου Γ και Φ είναι κατάλληλα επιλεγμένες σταθερές. Τουλάχιστον αυτή η μορφή ικανοποιεί τους στοχαστικούς περιορισμούς προσέγγισης.

Αριθμός βημάτων εκμάθησης

Αφού η εκμάθηση είναι μια στοχαστική διαδικασία, η τελική στατιστική ακρίβεια της χαρτογράφησης εξαρτάται από τον αριθμό των βημάτων στην τελική φάση σύγκλισης, η οποία πρέπει να είναι εύλογα μακρά. Δεν υπάρχει κανένας τρόπος να παρακαμφθεί αυτή η απαίτηση. Μια εμπειρική μέθοδος είναι ότι, για καλή στατιστική ακρίβεια, ο αριθμός βημάτων πρέπει να είναι τουλάχιστον 500 φορές ο αριθμός των στοιχείων του δικτύου. Χαρακτηριστικά, μέχρι 100000 βήματα έχουν χρησιμοποιηθεί στις προσομοιώσεις, αλλά για ταχύτερη εκμάθηση 10000 βήματα ή και ακόμα λιγότερα μπορούν μερικές φορές να είναι αρκετά. Επίσης, εάν σχετικά λίγα δείγματα είναι διαθέσιμα, πρέπει να ανακυκλωθούν για τον επιθυμητό αριθμό βημάτων. Διάφορες εναλλακτικές λύσεις υπάρχουν προς τούτο: τα δείγματα μπορούν να εφαρμοστούν κυκλικά ή σε μια τυχαία μεταβαλλόμενη διάταξή, ή να επιλεγούν τυχαία από το βασικό σύνολο (bootstrap learning). Έχει αποδειχθεί στην πράξη ότι η κυκλική εφαρμογή των δειγμάτων δεν είναι καταφανώς χειρότερη από τις άλλες, από μαθηματική απόψεως καλύτερα δικαιολογήσιμες, μεθόδους.

Βελτίωση των σπάνιων περιπτώσεων

Σε πολλά πρακτικά προβλήματα σημαντικές περιπτώσεις (δεδομένων εισόδου) μπορεί να εμφανιστούν με τη μικρή στατιστική συχνότητα, οπότε δεν είναι ικανά να καταλάβουν κάποιο τμήμα στο χώρο απεικόνισης (SOHMMM). Επομένως, τέτοιες σημαντικές

περιπτώσεις μπορούν να ενισχυθούν στην εκμάθηση αυθαίρετα, με τη χρήση μιας υψηλότερης τιμής για το α ή το hci για αυτά τα δείγματα, ή την επανάληψη αυτών των δειγμάτων σε μια τυχαία διάταξη για έναν ικανοποιητικό αριθμό φορών κατά τη διάρκεια της διαδικασίας εκμάθησης. Είναι φανερό πως ο προσδιορισμός της ζητούμενης ενίσχυσης κάποιων δειγμάτων κατά την εκμάθηση πρέπει να γίνεται σε συνεργασία με τους τελικούς χρήστες αυτών των χαρτών.

Επιβάλλοντας απεικονίσεις δεδομένων σε μια επιθυμητή θέση στο χάρτη

Μερικές φορές, ειδικά όταν το SOM χρησιμοποιείται για να ελέγξει τα πειραματικά δεδομένα, μπορεί να είναι επιθυμητό να χαρτογραφηθούν τα «κανονικά» δεδομένα σε μια καθορισμένη θέση (πχ, στη μέση) του χάρτη. Προκειμένου να "αναγκαστούν" να τοποθετηθούν τα συγκεκριμένα στοιχεία στις επιθυμητές θέσεις, είναι ενδεδειγμένο να χρησιμοποιηθούν τα αντίγραφά τους για τις αρχικές τιμές των διανυσμάτων αναφοράς σε αυτές τις θέσεις, και να κρατηθεί ο ρυθμός εκμάθησης α χαμηλός για αυτές τις θέσεις-κόμβους κατά τη διάρκεια της ενημέρωσής τους.

Διφορούμενα σύμβολα

Επειδή οι ακολουθιακές τεχνικές δεν είναι τέλειες, παρατηρούνται περιστασιακά διφορούμενα σύμβολα. Για παράδειγμα, το X αντιπροσωπεύει το A ή το C ή το G ή το T στις ακολουθίες DNA, και το B αντιπροσωπεύει asparagines ή ασπαρτικά οξέα στις πρωτεϊνικές ακολουθίες. Τέτοια σύμβολα μπορούν εύκολα να αντιμετωπιστούν με διάφορους τρόπους με τα HMMs. Στις αναζητήσεις βάσεων δεδομένων, είναι συνετή πρακτική να χρησιμοποιηθεί η προσέγγιση «benefit of the doubt», στην οποία ένα διφορούμενο σύμβολο αντικαθίσταται από την πλέον πιθανή εναλλακτική λύση του στον υπολογισμό της ακολουθιών likelihoods και των μονοπατιών Viterbi. Πρόσθετη προσοχή πρέπει να χρησιμοποιηθεί με τις ακολουθίες που έχουν ένα ασυνήθιστα μεγάλο μέρος διφορούμενων συμβόλων, δεδομένου ότι αυτές είναι πιθανό να παραγάγουν ψευδείς θετικές αποκρίσεις.

Σημασία του Emission Probability Matrix

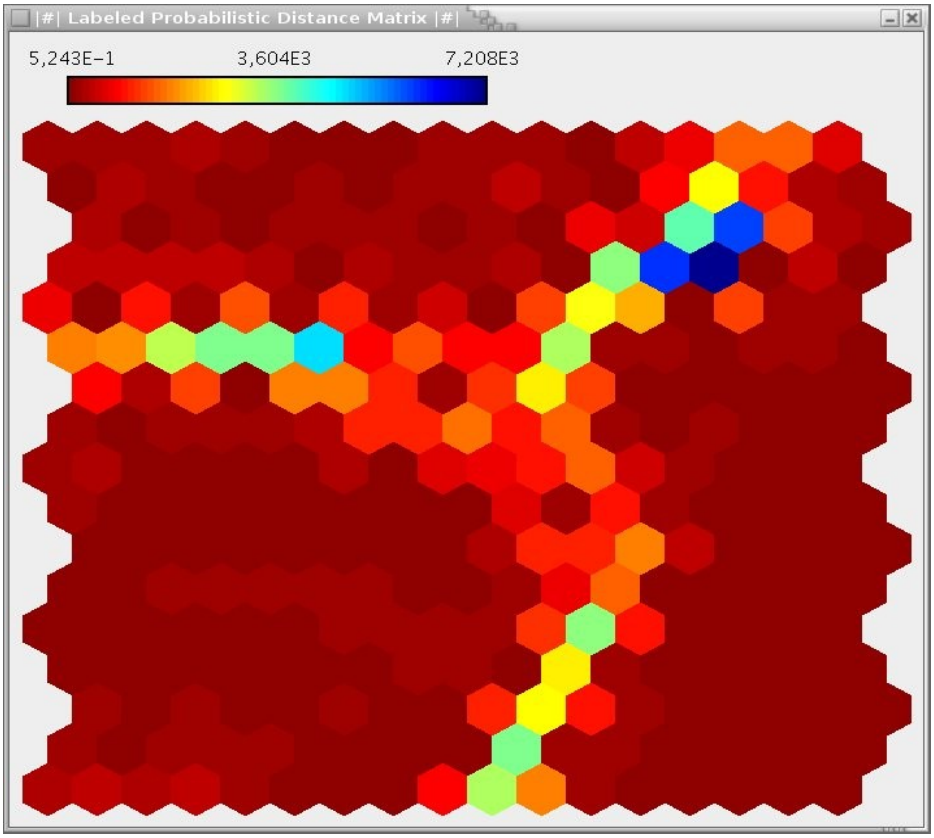
Μια άλλη σημαντική ιδιότητα του μέτρου απόστασης είναι ότι οι αποκλίσεις στη μήτρα B πιθανότητας εκπομπής δίνουν, γενικά, μεγαλύτερα αποτελέσματα απόστασης από τις παρόμοιες αποκλίσεις στη μήτρα A, πιθανότητας μετάβασης. Κατά συνέπεια, η μήτρα B εμφανίζεται να είναι αριθμητικά σημαντικότερη από τη μήτρα A στον προσδιορισμό ενός HMM.

Αποτίμηση

Όταν ένας ικανοποιητικός αριθμός ακολουθιών O_d έχει παρουσιαστεί και τα HMMs λ_e έχουν συγκλίνει σε σχεδόν στάσιμες τιμές, το επόμενο βήμα είναι αποτίμηση(calibration) του χάρτη, προκειμένου να βρεθούν οι εικόνες των διαφορετικών διαφορετικών ακολουθιών παρατήρησης σε αυτόν. Στις πρακτικές εφαρμογές για τις οποίες τέτοιοι χάρτες χρησιμοποιούνται, μπορεί να είναι αυτονόητο πώς ένα ιδιαίτερο σύνολο δεδομένων εισόδου οφείλει να ερμηνευθεί και να ονομαστεί. Με την εισαγωγή διάφορων μη αυτόματα αναλυθέντων σύνολων δεδομένων, βλέποντας που επιτυγχάνονται οι καλύτερες αντιστοιχίες στο χάρτη και ονομάζοντας τα στοιχεία του χάρτη αναλόγως, ο χάρτης αποτιμάται. Δεδομένου ότι αυτή η χαρτογράφηση υποτίθεται συνεχής κατά μήκος μιας υποθετικής «ελαστικής επιφάνειας», οι άγνωστες ακολουθίες παρατήρησης προσεγγίζονται από τις πιο κοντινές ακολουθίες παρατήρησης αναφοράς, όπως στον κβαντισμό διανυσμάτων.

Probabilistic Distance Matrix (PDM)

Το SOHMMM συνδυάζει συνδυάζει τις μεθόδους της μη γραμμικής προβολής κα της ομαδοποίησης σε έναν διατεταγμένο γράφο κβαντισμού διανυσμάτων. Στη χαρτογράφηση που παράγει, ακόμα και μετά την αποτίμηση, είναι δύσκολο να η ανιχνευθούν τα όρια μεταξύ των ενδεχόμενων συστάδων. Ένας γραφικός τρόπος αναπαράστασης, η αποκαλούμενη πιθανοτική μήτρα απόστασης (PDM), έχει αναπτυχθεί [8] για να επεξηγήσει την ομαδοποίηση των HMMs αναφοράς στο SOHMMM. Στην πραγματικότητα, ο PDM είναι μια επέκταση στην έννοια του U-Matrix. Στο [8] προτείνεται μια μέθοδος στην οποία οι πιθανοτικές αποστάσεις (ή πιθανοτικές ομοιότητες ή μέτρα απόκλισης) μεταξύ γειτονικών HMMs αντιπροσωπεύονται από σκιές σε μια γκρίζα κλίμακα (ή και ψευδοχρωματικές κλίμακες δύναται να χρησιμοποιηθούν). Εάν η πιθανοτική απόσταση γειτονικών HMMs είναι μικρή, μια ελαφριά σκιά χρησιμοποιείται και το αντίθετο, σκοτεινές σκιές αντιπροσωπεύουν μεγάλες πιθανοτικές αποστάσεις. Ένα τοπίο συστάδων που διαμορφώνεται στη συνέχεια στο SOHMMM οπτικοποιεί σαφώς την μη επιβλεπόμενη ομαδοποίηση. Στην εικόνα 4.3 απεικονίζεται ένα τέτοιο παράδειγμα



Εικόνα 4.3 Probabilistic Distance Matrix

4.4 Αλγόριθμος εκμάθησης για το SOHMMM

Προηγουμένως, έχει παρουσιαστεί και έχει αναλυθεί εξαντλητικά η βασική θεωρία του SOHMMM υπό το πρίσμα διάφορων ζητημάτων εφαρμογής. Είναι λογικό ότι αυτή η συζήτηση θα μπορούσε να δώσει αφορμή για διάφορες πιθανές παραλλαγές στον αλγόριθμο, τη μορφή και την αρχιτεκτονική του SOHMMM. Κατωτέρω, δίνεται μια γενική, και στο μεταξύ, αναλυτική και πλήρης περιγραφή του αλγόριθμου εκμάθησης του SOHMMM.

Συμβολίζουμε το σύνολο των ακολουθιών παρατηρήσεων ως

$$\boldsymbol{OS} = \{\boldsymbol{O}^{(1)}, \boldsymbol{O}^{(2)}, \ldots, \boldsymbol{O}^{(D)}\}$$

όπου $O^{(d)} = o_1 o_2 \ldots o_{Td}$ είναι η d-στή ακολουθία παρατήρησης, κάθε ακολουθία o_{td} λαμβάνει μια τιμή από το αλφάβητο $F = \{f_1, f_2, \ldots, f_G\}$, και T_d είναι ο αριθμός των παρατηρήσεων στην ακολουθία $O^{(d)}$. Επίσης, $\lambda = (A, B, \pi, W, R, U)$ είναι μια κλάση HMMs με τα ακόλουθα χαρακτηριστικά:

(i) τα αντίστοιχα σύμβολα παρατηρήσεων που προσδιορίζονται ($V = \{v_1, v_2, \ldots, v_M\}$) αποτελούν το ίδιο αλφάβητο με το F (ii) δύο μεμονωμένα HMMs λε, λe έχει ίσες πληθικότητες των αντίστοιχων χώρων καταστάσεων (συγκεκριμένα $N^{(\varepsilon)} = N^{(e)}$) και έχουν,

γενικά, διαφορετικές τιμές στοιχείων των πιθανοτικών μητρών A και B, των αρχικών κατανομών πιθανότητας π και των μητρών W, R και U (iii) ο αριθμός καταστάσεων N σε ολόκληρη την κλάση λ είναι καθορισμένος αυθαίρετα ή αντιστοιχεί κατά προσέγγιση στον μέσο αριθμό παρατηρήσεων μεταξύ των στοιχείων του συνόλου ακολουθιών παρατηρήσεων.

Επίσης η κανονικοποίηση των μπροστινών και οπίσθιων μεταβλητών επιτυγχάνεται με τη χρήση των λημμάτων 7 – 9 (βλ. παράρτημα B).

Το SOHMMM αποτελείται από E HMMs, τα οποία είναι στιγμιότυπα της κλάσης λ. Αυτά τα HMMs αντιστοιχίζονται στους κόμβους ενός δισδιάστατου εξαγωνικού πλέγματος. Υπάρχουν E1 HMMs ανά γραμμή και E2 HMMs ανά στήλη πλέγματος. Τα E1 και E2 επιλέγονται κατάλληλα ώστε να ανταποκριθούν σε δύο προδιαγραφές. Κατ' αρχάς, E1*E2 = E (για προφανείς λόγους) και δεύτερον οι τιμές των E1 και E2 πρέπει να επιβάλουν μια επιμήκη μορφή του χάρτη (με άλλα λόγια οι άκρες του πλέγματος γίνονται ορθογώνιες).

Κατά τη διάρκεια της φάσης διάταξης του αλγορίθμου εκμάθησης SOHMMM χρησιμοποιείται μια γκαουσιανή συνάρτηση γειτονιάς. Η τυπική απόκλιση της συνάρτησης αρχίζει με μια τιμή του ίδιου μεγέθους με τη μεγαλύτερη διάσταση του πλέγματος και μειώνεται σε μια μονάδα χάρτη στο τέλος της φάσης αυτής. Ο ρυθμός εκμάθησης μειώνεται γραμμικά, αρχίζοντας από υψηλές τιμές (κοντά στην μονάδα) και στο τέλος αποκτά τιμές της τάξης του 0.01.

Κατόπιν, κατά τη διάρκεια της φάσης tuning το πλάτος της γειτονιάς παραμένει σταθερό και το NBc περιέχει τους πλησιέστερους γείτονες του HMM c. Η ακολουθία {η(y)} καθορίζεται από μια συνάρτηση αντιστρόφως ανάλογη προς το Y, λαμβάνοντας τιμές στο διάστημα [0.01, 0.0001].

Παρακάτω δίνεται ο ψευδοκώδικας για το SOHMMM:

/*initialization*/
1] for *e* = 1 to *E*

2] $w_{ij}^{(e)} \leftarrow$ **uniformDistribution(), 1≤ *i*≤ *N*, 1≤ *j*≤ *N*;**

3] $r_{jt}^{(e)} \leftarrow$ **uniformDistribution(), 1≤ *j*≤ *N*, 1≤ *t*≤ *M*;**

4] $u_j^{(e)} \leftarrow$ **uniformDistribution(), 1≤ *j*≤ *N*;**

5] $a_{ij}^{(e)} \leftarrow e^{w_{ij}^{(e)}} / \sum_{l=1}^{N} e^{w_{il}^{(e)}}$ **, 1≤ *i*≤ *N*, 1≤ *j*≤ *N*;**

6] $b_j^{(e)}(t) \leftarrow e^{r_{jt}^{(e)}} / \sum_{l=1}^{M} e^{r_{jl}^{(e)}}$ **, 1≤ *j*≤ *N*, 1≤ *t*≤ *M*;**

7] $\pi_j^{(e)} \leftarrow e^{u_j^{(e)}} / \sum_{l=1}^{N} e^{u_l^{(e)}}$ **, 1≤ *j*≤ *N*;**

end for

/*ordering Phase*/
8] $\eta(y) \leftarrow \eta_{ordering}(y)$**;**
9] for *y* = 1 to *orderingSteps*

10] $O \leftarrow$ **randomCyclicSelection(*OS*);**

11] $c \leftarrow \arg\min_{e} \{-\log P(O \mid \lambda_e)\}$ **;**

12] for *e* = 1 to *E*

13] $w_{ij}^{(e)} \leftarrow$

$$w_{ij}^{(e)} + \eta(y)\exp\left(-\frac{\|\delta_c - \delta_e\|^2}{2\sigma^2(y)}\right)\left[\frac{a_{ij}}{P(O|\lambda)}\sum_{l=1}^{T-1}\left[\alpha_l(i)b_j(o_{l+1})\beta_{l+1}(j) - \alpha_l(i)\beta_l(i)\right]\Big|_{\lambda_e}\right]$$

$1 \le i \le N,\ 1 \le j \le N;$

14] $r_{jt}^{(e)} \leftarrow$

$$r_{jt}^{(e)} + \eta(y)\exp\left(-\frac{\|\delta_c - \delta_e\|^2}{2\sigma^2(y)}\right)\left[\frac{1}{P(O|\lambda)}\sum_{l=1}^{T}\left[I\{o_l = t|\lambda\}\alpha_l(j)\beta_l(j) - b_j(t)\alpha_l(j)\beta_l(j)\right]\Big|_{\lambda_e}\right]$$

$1 \le j \le N,\ 1 \le t \le M;$

15] $u_j^{(e)} \leftarrow$

$$u_j^{(e)} + \eta(y)\exp\left(-\frac{\|\delta_c - \delta_e\|^2}{2\sigma^2(y)}\right)\left[\pi_j\left[\frac{b_j(o_1)\beta_1(j)}{P(O|\lambda)} - 1\right]\Big|_{\lambda_e}\right]$$

$1 \le j \le N;$

16] $a_{ij}^{(e)} \leftarrow e^{w_{ij}^{(e)}} / \sum_{l=1}^{N} e^{w_{il}^{(e)}}$, $1 \le i \le N,\ 1 \le j \le N;$

17] $b_j^{(e)}(t) \leftarrow e^{r_{jt}^{(e)}} / \sum_{l=1}^{M} e^{r_{jl}^{(e)}}$, $1 \le j \le N,\ 1 \le t \le M;$

18] $\pi_j^{(e)} \leftarrow e^{u_j^{(e)}} / \sum_{l=1}^{N} e^{u_l^{(e)}}$, $1 \le j \le N;$

end for

end for

/*tuning phase*/

19] $\eta(y) \leftarrow \eta_{tuning}(y)$;

20] for y = 1 to *tuningSteps*

21] $O \leftarrow$ randomCyclicSelection(OS);

22] $c \leftarrow \arg\min_{e}\left\{-\log P(O|\lambda_e)\right\}$;

23] $NB_c \leftarrow$ nearestNeighbors(c);

24] for e = 1 to E

25] if $e \in NB_c$

26] $w_{ij}^{(e)} \leftarrow$

$$w_{ij}^{(e)} + \eta(y)\left[\frac{a_{ij}}{P(O|\lambda)}\sum_{l=1}^{T-1}\left[\alpha_l(i)b_j(o_{l+1})\beta_{l+1}(j) - \alpha_l(i)\beta_l(i)\right]\Big|_{\lambda_e}\right]$$

$1 \le i \le N,\ 1 \le j \le N;$

27] $r_{jt}^{(e)} \leftarrow$

$$r_{jt}^{(e)} + \eta(y)\left[\frac{1}{P(O|\lambda)}\sum_{l=1}^{T}\left[I\{o_l = t|\lambda\}\alpha_l(j)\beta_l(j) - b_j(t)\alpha_l(j)\beta_l(j)\right]\Big|_{\lambda_e}\right]$$

$1 \le j \le N,\ 1 \le t \le M;$

28] $u_j^{(e)} \leftarrow u_j^{(e)} + \eta(y)\left[\pi_j\left[\frac{b_j(o_1)\beta_1(j)}{P(O|\lambda)} - 1\right]\Big|_{\lambda_e}\right]$

$1 \le j \le N;$

29] $a_{ij}^{(e)} \leftarrow e^{w_{ij}^{(e)}} / \sum_{l=1}^{N} e^{w_{il}^{(e)}}$, $1 \leq i \leq N$, $1 \leq j \leq N$;

30] $b_j^{(e)}(t) \leftarrow e^{r_{jt}^{(e)}} / \sum_{l=1}^{M} e^{r_{jl}^{(e)}}$, $1 \leq j \leq N$, $1 \leq t \leq M$;

31] $\pi_j^{(e)} \leftarrow e^{u_j^{(e)}} / \sum_{l=1}^{N} e^{u_l^{(e)}}$, $1 \leq j \leq N$;

end if

end for

end for

Κεφάλαιο 5
Μέτρα και κριτήρια της ποιότητας για το υβριδικό νευρωνικό δίκτυο SOHMMM

5.1 Εισαγωγή

Σε αυτό το κεφάλαιο παρουσιάζεται και αναλύεται μια πλειάδα από κριτήρια και μέτρα που ποσοτικοποιούν την ποιότητα της οργάνωσης και της απεικόνισης (mapping) για το υβριδικό νευρωνικό δίκτυο SOHMMM. Ένας μεγάλος αριθμός ερευνητών έχει ασχοληθεί με την ανάπτυξη αυτών των κριτηρίων, που επιτρέπουν να αποφανθούμε για το πόσο καλή είναι μια απεικόνιση ή πόσο καλά έχει οργανωθεί μια χαρτογράφηση δεδομένων, και ειδικά για τον αυτο-οργανούμενο χάρτη (SOM) υφίσταται μια αρκετά εκτενής βιβλιογραφία. Η συμβολή αυτής της εργασίας συνίσταται στο να μελετηθούν και να αναλυθούν τα υπάρχοντα αυτά κριτήρια και στη συνέχεια, με βάση αυτά, να αναπτυχθούν νέα ειδικά προσαρμοσμένα για το SOHMMM ικανά να ποσοτικοποιήσουν την διαδικασία εκπαίδευσης και επιλογής παραμέτρων και να οδηγήσουν σε μια χρήσιμη εξαγωγή συμπερασμάτων για την ποιότητα της οργάνωσης του χάρτη. Στις επόμενες ενότητες αρχικά γίνεται μια εισαγωγή σε κάποιες χρήσιμες έννοιες όσον αφορά τα μέτρα οργάνωσης, ενώ στη συνέχεια παρουσιάζονται τα κριτήρια και αναφερόμαστε εκτενώς στο πώς προσαρμόστηκε το καθένα κατάλληλα. Γίνεται επίσης μια αναφορά σε κριτήρια που μελετήθηκαν, αλλά δεν χρησιμοποιήθηκαν τελικά, με παράλληλη επεξήγηση των λόγων που δεν κατέστη δυνατό αυτό.

5.2 Γενικές έννοιες και ορισμοί

Αν και φαίνεται ότι υφίσταται μια κοινή διαισθητική αντίληψη της έννοιας της τοπογραφίας ενός χάρτη, στην πράξη έχει αποδειχθεί δύσκολο να περιγραφεί αυτή η έννοια με έναν σαφή και αυστηρό από μαθηματικής απόψεως τρόπο. Με έναν άτυπο ορισμό, η τοπογραφία ενός χάρτη σημαίνει την απεικόνιση όμοιων σημείων δεδομένων σε κοντινές θέσεις στο πλέγμα του χάρτη.

Υπάρχουν διάφορες επιλογές για τον ορισμό μιας χαρτογράφησης που διατηρεί ιδανικά τη γειτνίαση των δεδομένων από ένα χώρο σε έναν άλλο. Η αυστηρότερη είναι να απαιτήσουμε η χαρτογράφηση να διατηρεί τις ομοιότητες, δηλαδή για κάθε ζευγάρι σημείων σε ένα χώρο, η ομοιότητά τους να είναι ίση με την ομοιότητα των απεικονίσεών τους στον άλλο χώρο. Μια λιγότερο αυστηρή είναι οι τιμές ομοιότητας μεταξύ των ζευγών των σημείων να συσχετίζονται τέλεια. Λιγότερο αυστηρή ακόμα είναι η επιλογή η χαρτογράφηση να πρέπει μόνο να διατηρήσει την διάταξη ομοιότητας, δηλαδή αντί να συγκρίνουμε τις απόλυτες τιμές της ομοιότητας μεταξύ των ζευγών των σημείων σε ένα χώρο και της ομοιότητας μεταξύ των απεικονίσεων τους σε έναν άλλο, μας ενδιαφέρει μόνο η σχετική διάταξη των ομοιοτήτων μέσα στα δύο σύνολα να είναι η ίδια. Εάν οι τιμές ομοιότητας στο χώρο εξόδου σχεδιάζονται σε σχέση με τις τιμές ομοιότητας στο χώρο εισόδου για έναν συγκεκριμένο χάρτη, το πρώτο κριτήριο προδιαγράφει ότι όλα τα σημεία πρέπει να βρίσκονται σε μια ευθεία γραμμή,που σχηματίζει γωνία 45 μοιρών με τον άξονα των x, το δεύτερο ότι τα σημεία βρίσκονται σε μια ευθεία γραμμή αυθαίρετης γωνίας, και το τρίτο ότι τα σημεία βρίσκονται σε μια γραμμή που δεν είναι απαραιτήτως ευθεία, αλλά αυξάνεται μονοτονικά. Είναι σημαντικό να είναι σαφές ποιος από αυτούς τους τρεις στόχους (ή ίσως κάποιος τέταρτος) υπονοείται από κάποια συγκεκριμένη φιλοσοφία χαρτογράφησης. Διαφορετικοί στόχοι θα είναι κατάλληλοι για διαφορετικές εφαρμογές.

Εντούτοις, στις περισσότερες πρακτικές εφαρμογές καμία από αυτές τις επιλογές ιδανικού χάρτη δεν θα είναι επιτεύξιμη, και ένα μέτρο απόκλισης απαιτείται που να αξιολογεί το βαθμό στον οποίο η τελειότητα έχει επιτευχθεί. Λαμβάνοντας υπόψη έναν ορισμό της

τελειότητας υπάρχουν πολλοί τρόποι να μετρηθεί η απόκλιση, κάθε ένας από τους οποίους συγκρίνει τους μη-ιδανικούς χάρτες με διαφορετικό τρόπο.

5.2.1 Έννοιες για τα μέτρα οργάνωσης

Χώρος εισόδου και χώρος εξόδου. Εν γένει, θα θεωρήσουμε δύο χώρους, έναν χώρο εισόδου V και έναν χώρο εξόδου A στον οποίο απεικονίζονται τα δείγματα από τον V. Ειδικά για την περίπτωση του SOHMMM, πρέπει να τονιστεί πως το V δεν μπορεί να ονομαστεί τυπικά χώρος καθώς τα σήματα εισόδου μπορούν να είναι ακολουθίες κάποιου αλφαβήτου, δείγματα φωνής, κλπ και επομένως δεν είναι δυνατός ο ορισμός της έννοιας της απόστασης (τουλάχιστον με την κλασική έννοια που υφίσταται στο R^n) και δεν ισχύει προφανώς η τριγωνική ανισότητα. Για λόγους απλότητας πάντως θα χρησιμοποιούμε τον συμβολισμό V για το σύνολο των δεδομένων εισόδου. Επίσης, σε ότι αφορά τον χώρο A, πρέπει να τονιστεί πως αναφερόμαστε σε έναν διακριτό συνήθως χώρο, χαμηλής διάστασης (συνήθως το πολύ τρισδιάστατο).

Μέτρα πρώτης και δεύτερης τάξης. Στην ανάλυσή μας δεν λήφθηκαν υπόψιν μόνο "καθαρά" μέτρα οργάνωσης που ποσοτικοποιούν την οργανωτική δομή της χαρτογράφησης-απεικόνισης αλλά συμπεριλήφθηκαν και μέτρα κβαντισμού ή παραμόρφωσης και μετάδοσης πληροφορίας (εντροπία). Τα πρώτα μπορούμε να τα ονομάσουμε πρώτης τάξης και τα τελευταία δεύτερης τάξης. Μιλώντας για μέτρα πρώτης τάξης, εννοούμε ότι προκύπτουν από τον συνδυασμό τιμών (πχ συχνότητα ενεργοποίησης ή ένταση) που λαμβάνονται για μεμονωμένους νευρώνες, ενώ τα μέτρα δεύτερης τάξης έχουν τιμές που προκύπτουν από συνδυασμό τιμών (πχ ομοιότητα ή απόσταση) που λαμβάνονται για ζευγάρια νευρώνων.

Δεδομενοκεντρική προσέγγιση. Οι προσεγγίσεις για την μέτρηση της οργάνωσης μπορούν να ταξινομηθούν ανάλογα με το αν λαμβάνεται υπόψιν η δομή των δεδομένων που θα απεικονιστούν στους υπολογισμούς ή όχι. Με δεδομενοκεντρικά μέτρα, δομικά ισοδύναμες χαρτογραφήσεις μπορούν να έχουν εντελώς διαφορετικές τιμές μέτρων για τον ίδιο τύπο μέτρου, ανάλογα με την δομή των δεδομένων.

Κατηγορήματα, μέτρα και παράμετροι τάξης. Υπάρχουν διάφορες έννοιες της διατήρησης τοπολογίας που απλά διακρίνουν τις περιπτώσεις όπου η απεικόνιση είναι διατεταγμένη ή όχι. Εντούτοις, αυτού του είδους ο χαρακτηρισμός έχει διαφορετική ποιότητα από μέτρα που αντιστοιχίζουν μια αριθμητική τιμή στην μη διατεταγμένη περίπτωση. Η πρώτη περίπτωση μπορεί να ιδωθεί ως ένα λογικό κατηγόρημα που ορίζει την έννοια μια απεικόνιση να είναι διατεταγμένη (ή όχι) με έναν στενά μαθηματικό τρόπο. Όταν για παράδειγμα το κατηγόρημα διάταξης δεν ικανοποιείται, τότε ο βαθμός αταξίας, δηλαδή η απόκλιση από την ταξινομημένη περίπτωση, μπορεί να ποσοτικοποιηθεί περαιτέρω. Για αυτή την ποσοτικοποίηση υπάρχει ένας μεγάλος βαθμός αυθαιρεσίας.

Ένα μέτρο οργάνωσης μπορεί να προκύψει έτσι ώστε να λαμβάνει κάποια ακρότατη (μέγιστη ή ελάχιστη) τιμή για την ταξινομημένη περίπτωση και να αποκλίνει από αυτή ανάλογα με το πόσο "μακριά" από την ταξινομημένη μορφή βρίσκεται η απεικόνιση. Τέτοιου είδους μέτρα συναντούμε στους [2] και [3].

Η φιλοσοφία είναι παρόμοια με αυτή των παραμέτρων τάξης που συναντούνται στη θερμοδυναμική. Οι παράμετροι αυτοί χρησιμοποιούνται στη φυσική ώστε να διακρίνουμε διαφορετικούς τύπους καταστάσεων ισορροπίας σε φυσικά συστήματα. Συνήθως, επιλέγονται έτσι ώστε η τιμή τους να εξαφανίζεται σε καταστάσεις με υψηλότερη συμμετρία

κατανομής και να αποκλίνει από το 0 όταν η συμμετρία της κατανομής κατάστασης παύει να υφίσταται. Η προσέγγιση των παραμέτρων διάταξης έχει χρησιμοποιηθεί άμεσα για την μελέτη των SOMs και άλλων συναφών μοντέλων [4] [5] [6].

Τοπολογία και γεωμετρία. Η οργάνωση μιας απεικόνισης συχνά αντιμετωπίζεται ως ισοδύναμη με την διατήρηση της τοπολογίας. Σε μια πιο αυστηρή διατύπωση, τα περισσότερα μέτρα δεν μετρούν την διατήρηση της τοπολογίας αλλά χρησιμοποιούν μια μίξη από ιδιότητες ομοιότητας, τοπολογικές, μετρικές ή ακόμα και γεωμετρικές ιδιότητες των χώρων που απεικονίζονται. Τα μέτρα μπορεί να ενσωματώνουν τα ίδια μετρικές ομοιότητας ή απλά την σχετική τους διάταξη. Παρακάτω δίνεται μια σύντομη ταξινόμηση ανάλογα με τις διάφορες προσεγγίσεις.

"Καθαρά" μέτρα τοπολογίας. Πολύ λίγα μέτρα οργάνωσης έχουν διατυπωθεί με όρους καθαρής τοπολογίας. Ένα από τα προβλήματα είναι να βρεθεί τέτοια διατύπωση ώστε ένας από τους χώρους που εμπλέκονται σε μια χαρτογράφηση να είναι συνήθως διακριτός. Για αυτούς τους χώρους η κανονική τοπολογία είναι η διακριτή, πράγμα καθόλου χρήσιμο αφού ουσιαστικά σημαίνει πλήρη απουσία δομής.

Μια προσέγγιση αναλυτικής επίλυσης του προβλήματος παρουσιάστηκε στο [2] με χρήση μιας συλλογής από προσημασμένους διακριτούς χώρους στους οποίους η τοπολογία ορίζεται σύμφωνα με το αντίστοιχο προσημασμένο στοιχείο. Αυτή η προσέγγιση απαιτεί μια επέκταση της κανονικής έννοιας της τοπολογίας στον διακριτό χώρο. Επιπρόσθετα, οι τοπολογικές δομές που θεωρούνται σε αυτό το κείμενο [2] συνάγονται από δομές μετρικών.

Στα [7], [8], [9] διερευνήθηκαν ισοπίθανες μέθοδοι απεικόνισης, βασισμένες σε θεωρήσεις μεγιστοποίησης της μετάδοσης πληροφορίας από τις απεικονίσεις. Σε αυτό το σημείο, θα ήταν καλό να αναφερθεί πως η δομή που εισάγεται στις παραπάνω δημοσιεύσεις όσον αφορά τον διακριτό χώρο μπορεί να γίνει αντιληπτή ως ένας complex, μια δομή γνωστή από την αλγεβρική τοπολογία. Ένας complex μπορεί να θεωρηθεί ως μια γενίκευση της έννοιας ενός γράφου. Παρολ' αυτά, δεν απαιτείται κάποια παραγωγή δομών μετρικών για τον ορισμό του και επομένως η μέθοδος και τα μέτρα που παράγονται από αυτόν μπορούν να θεωρηθούν ως πραγματικά καθαρά τοπολογικές έννοιες.

Γεωμετρία. Πολύ λίγα μοντέλα χρησιμοποιούν γεωμετρικές ιδιότητες προκειμένου να ποσοτικοποιήσουν την οργάνωση. Στο [10] παρουσιάζεται μια προσέγγιση που υποθέτει τον χώρο εξόδου ως ορθογωνικό πλέγμα και σχετίζεται με τα μέτρα καμπυλότητας των χώρων Riemann.

Ομοιότητα, απόσταση και μετρικές. Τα ζεύγη των στοιχείων των χώρων εισόδου και εξόδου πρέπει να συγκριθούν ώστε να ποσοτικοποιηθεί η οργάνωση της χαρτογράφησης. Αυτό μπορεί να συμβεί σε επίπεδο γράφου, προσδιορίζοντας αν τα στοιχεία του χώρου είναι γειτονικά ή μη. Ένας δεύτερος, πιο γενικός τρόπος είναι να ορίσουμε ένα μέτρο στον χώρο που καθορίζει πόσο όμοια (ή ανόμοια) είναι τα στοιχεία του χώρου. Είναι πολύ σημαντικό να τονιστεί ότι ένα μέτρο ομοιότητας δεν είναι απαραίτητο να έχει μια μετρική ή γεωμετρική ερμηνεία, αντίθετα με μία μετρική που πρέπει να ικανοποιεί την τριγωνική ανισότητα. Η τριγωνική ανισότητα, εντούτοις, μπορεί να θεωρηθεί ως ένα είδος γεωμετρικής γενίκευσης της μεταβατικής ιδιότητας μιας σχέσης γραμμικής διάταξης.

Συνεπώς, είναι αρκετά φυσικό να συμπεριλάβουμε χώρους βασισμένους σε δομή γράφων, αφού ένας γράφος μπορεί να θεωρηθεί σαν μια τοπολογική αλλά και μετρική δομή. Εντούτοις, δεν είναι δυνατόν να εκμεταλλευθούμε πλήρως τον γράφο σαν καθαρά τοπολογική δομή στην περίπτωση περισσότερων της μιας διάστασης, όπου κανείς οφείλει να χρησιμοποιήσει δομές όπως οι τοπολογικοί complex. Πάντως, ο γράφος μπορεί να

χρησιμοποιηθεί πάντοτε ως μετρική δομή (και επομένως, μέσω κάποιας πχ συνάρτησης γειτονιάς ως μέτρο της ομοιότητας). Αυτές οι ιδιότητες κάνουν τον γράφο ένα πολύ χρήσιμο μοντέλο για τον χώρο εξόδου V και σε πολλές περιπτώσεις απεικόνισης, όπως στην περίπτωση του SOM ή του SOHMMM, ο χώρος εξόδου μοντελοποιείται ως γράφος.

Σε αυτό το σημείο, είναι καλό να υπενθυμίσουμε τις συνθήκες που ένα μέτρο απόστασης και μια μετρική πρέπει να καλύπτουν. Αν συμβολίσουμε με d το μέτρο και x, y δύο οποιαδήποτε στοιχεία του χώρου για τον οποίο ορίζεται αυτό πρέπει:

- $d(x, y) \geq 0$ (μη αρνητικό)
- $d(x, y) = 0$ αν και μόνο αν $x = y$ (ταύτιση)
- $d(x, y) = d(y, x)$ (συμμετρία)
- $d(x, z) \leq d(x, y) + d(y, z)$ (τριγωνική ανισότητα)

Μια συνάρτηση που καλύπτει τους τρεις πρώτους περιορισμούς καλείται ένα μέτρο απόστασης. Αν επιπρόσθετα καλύπτει και τον τέταρτο περιορισμό, καλείται μετρική.

Όσον αφορά τον ορισμό της ομοιότητας, αν και οι ορισμοί που δίνονται στη βιβλιογραφία ποικίλουν μπορούμε εν γένει να πούμε πως ένα μέτρο ομοιότητας S πρέπει να καλύπτει για οποιαδήποτε x, y τους εξής περιορισμούς:

- $S(x, y) \geq 0$ (μη αρνητικό)
- $S(x, y) \leq S(x, x)$ και $S(x, y) = S(x, x)$ ανν $x = y$

Πρόσθετοι περιορισμοί είναι δυνατόν να απαιτηθούν, όπως το S να είναι συμμετρικό ή να έχει πεπερασμένη τιμή, όμως αυτές οι ιδιότητες δεν μας απασχολούν στον ορισμό των μέτρων ομοιότητας για το SOHMMM.

Στην περίπτωση του SOHMMM το συνηθέστερο μέτρο απόστασης που χρησιμοποιήθηκε για τον χώρο εξόδου (απόσταση κόμβων) είναι η ευκλείδεια απόσταση ενώ για τη απόσταση στο V (δηλαδή μεταξύ των HMMs) χρησιμοποιήθηκε το πιθανοτικό μέτρο απόστασης (probabilistic distance measure) από την εργασία των [11] όπου ορίζεται η έννοια της απόστασης μεταξύ δύο Hidden Markov Models (βλέπε και ενότητα 3.6). Υπενθυμίζεται εδώ, ότι στο κλασσικό SOM καθώς και στις περισσότερες μεθόδους απεικόνισης – χαρτογράφησης (πχ Sammon mappings κλπ) το σύνηθες μέτρο ομοιότητας τόσο στον χώρο εξόδου όσο και στον χώρο εισόδου αποτελεί η ευκλείδεια απόσταση.

Loglikelihood. Σε κάποια μέτρα οργάνωσης έπρεπε να οριστεί ένα μέτρο ομοιότητας (ή ανομοιότητας) μεταξύ των δεδομένων εισόδου και των κόμβων αναφοράς. Ο ορισμός ενός τέτοιου μέτρου, που στην περίπτωση του SOM είναι εξαιρετικά απλός αφού τόσο τα δεδομένα εισόδου όσο και οι κόμβοι αναφοράς μοντελοποιούνται ως διανύσματα, είναι στην περίπτωση του SOHMMM μια αρκετά περίπλοκη διαδικασία, αφού τα δεδομένα εισόδου είναι ακολουθίες ενός αλφαβήτου και οι κόμβοι αναφοράς μοντελοποιούνται ως HMM. Εν τέλει, επιλέχθηκε ως μέτρο ανομοιότητας η αρνητική λογαριθμική πιθανότητα (negative loglikelihood). Αν και με αυστηρούς μαθηματικούς όρους δεν μπορούμε να ονομάσουμε το μέγεθος αυτό μέτρο απόστασης εφόσον οι ακολουθίες και τα HMM δεν ανήκουν σε κάποιον χώρο και προφανώς δεν υφίσταται η έννοια της συμμετρίας, εντούτοις το μέτρο αυτό καλύπτει τους δύο πρώτους περιορισμούς ενός μέτρου απόστασης. Αν λ είναι ένα HMM και O είναι μια ακολουθία ενός πεπερασμένου αλφαβήτου, τότε διαδοχικά:

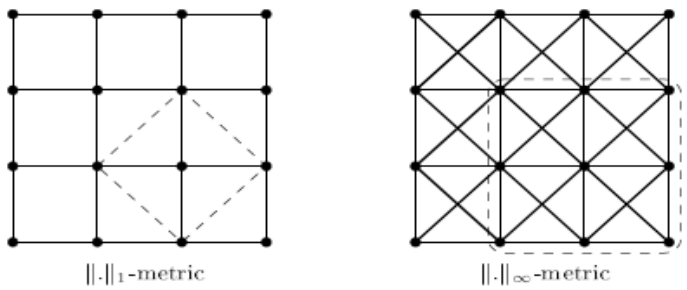
$$0 \leq P(O|\lambda) \leq 1 \Rightarrow -\infty \leq \log[P(O|\lambda)] \leq 0 \Rightarrow 0 \leq -\log[P(O|\lambda)] \leq \infty$$

ενώ ισχύει

- log [P(O|λ)] = 0 αν και μόνο αν το *λ* παράγει την ακολουθία *O*, δηλαδή η πιθανότητα εμφάνισης της ακολουθίας *O* από το *λ* είναι ίση με 1. Το παραπάνω είναι το ισοδύναμο του περιορισμού της ταύτισης στο κλασικό μέτρο απόστασης. Προφανώς, όσο πιο μικρή είναι η πιθανότητα να παράγει το *λ* την ακολουθία *O*, τόσο πιο μεγάλη είναι η τιμή της αρνητικής λογαριθμικής πιθανότητας, κατά αντιστοιχία με ένα μέτρο ανομοιότητας (απόστασης).

5.2.2 Μετρικές στον χώρο εξόδου και Τοπολογία

Ο περιορισμός σε μετρικές που προκύπτουν από γράφους στον χώρο εξόδου *A* επιτρέπει την πραγματοποίηση σημαντικών μετρικών δομών σε τυπικές δομές όπως ένα πλέγμα ν-διαστάσεων. Στην εικόνα 5.1 παρουσιάζονται δύο τέτοιες συγκεκριμένες μετρικές σε ένα δισδιάστατο πλέγμα νευρώνων. Οι μετρικές $\|.\|_1$ *και* $\|.\|_\infty$ μπορούν να πραγματοποιηθούν, όμως όχι η (ευκλείδεια) μετρική $\|.\|_2$.



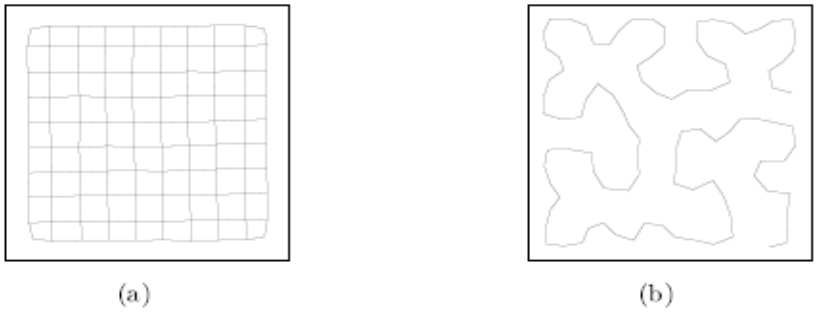
Εικόνα 5.1.Δύο διαφορετικές μετρικές δομές σε γράφους Kohonen.

Η προσέγγιση σύμφωνα με την οποία ο χώρος εξόδου είναι ένα μονοδιάστατο, δισδιάστατο ή υψηλότερης διάστασης πλέγμα (συνήθως ορθογωνικό) χρησιμοποιείται ευρέως στη βιβλιογραφία. Εν γένει, στην ανάλυσή μας θα προσπαθήσουμε να είμαστε όσο το δυνατόν λιγότερο περιοριστικοί σε ότι αφορά την διάσταση του πλέγματος.

5.2.3 Απαιτήσεις για ένα μέτρο οργάνωσης

Εξαιτίας του γεγονότος ότι δεν υφίσταται μια κανονική (τυπική) έννοια ή μέτρο της οργάνωσης ενός SOM και κατ' επέκταση και ενός SOHMMM, υπάρχει ένας αξιοσημείωτος βαθμός αυθαιρεσίας (βαθμών ελευθερίας) για την επιλογή ενός τέτοιου μέτρου. Θα είναι επομένως απαραίτητο να προσδιοριστούν με ακρίβεια οι απαιτήσεις που περιγράφουν μια έννοια οργάνωσης σε ένα SOHMMM. Έστω $W_{A,V} \equiv W$ το σύνολο των χαρτών με χώρο εξόδου *A* και σύνολο εισόδου *V*. Θέλουμε να ορίσουμε άτυπα δύο ιδιότητες που θα επιτρέπουν σε μια συνάρτηση μ: $W_{A,V} \rightarrow R$ να καλείται μέτρο οργάνωσης:

1. Πρέπει να ποσοτικοποιεί την διαδικασία της αυτο-οργάνωσης κατά τη διάρκεια της εκπαίδευσης, δηλαδή η τιμή της να αυξάνεται ή μειώνεται μονοτονικά κατά μέσο όρο.
2. Πρέπει να μετρά την ποιότητα της ενσωμάτωσης του *A* στον χώρο των δεδομένων, για παράδειγμα η χαρτογράφηση της εικόνας 5.2(α) πρέπει να έχει μια καλύτερη τιμή από την 5.2(β)



Εικόνα 5.2 δυο διαφορετικοί χάρτες Kohonen με διαφορετική ποιότητα απεικόνισης

Η πρώτη συνθήκη είναι ενδιαφέρουσα από μαθηματικής απόψεως και για την κατανόηση της διαδικασίας οργάνωσης. Η δεύτερη είναι περισσότερο σημαντική σε εφαρμογές που καλούνται να υπολογίσουν την ποιότητα της αφαίρεσης – περιγραφής ενός χώρου δεδομένων υψηλής διάστασης από έναν χάρτη απεικόνισης (SOM ή SOHMMM).

Εντούτοις, ένα μέτρο οργάνωσης δεν χρειάζεται να ικανοποιεί και τις δύο ιδιότητες. Μάλιστα, δεν είναι καν ξεκάθαρο ότι μπορεί να βρεθεί ένα μέτρο μ που να ικανοποιεί έστω το πρώτο κριτήριο. Η δεύτερη ιδιότητα δεν είναι ένας τόσο αυστηρός περιορισμός όσο ο πρώτος αφού διαφορετικά μέτρα θα έχουν κατά πάσα πιθανότητα διαφορετικές εκτιμήσεις για μη ξεκάθαρες περιπτώσεις απεικονίσεων.

Αφού η πρώτη ιδιότητα μπορεί να αποκτήσει μια συγκεκριμένη μορφή σχετικά εύκολα είναι προτιμότερο να επιλέξουμε αρχικά ένα μέτρο οργάνωσης που θα ικανοποιεί αυτή την ιδιότητα. Μπορούμε να απαιτήσουμε μια λιγότερο αυστηρή περίπτωση της 1 που θα απαιτούσε το μ να είναι απλώς μια συνάρτηση Liapunov, δηλαδή ότι το μ θα μειώνεται κατά μέσο όρο στη διάρκεια της εκπαίδευσης ή τυπικά

$$E(\mu(w(t+1))) \leqslant \mu(w(t))$$

για όλα τα *w(t)* που ανήκουν στο *W*, όπου συμβολίζουμε με *E* την μέση τιμή επί ενός τυχαίου σήματος εκπαίδευσης.

Ένα μέτρο οργάνωσης μπορεί να εκφράσει μια απόκλιση, δηλαδή η τιμή του μπορεί να μειώνεται όταν η οργάνωση βελτιώνεται κατά τη διάρκεια της εκπαίδευσης ή η τιμή του μπορεί να αυξάνεται κατά τη διάρκεια της διαδικασίας αυτο-οργάνωσης. Ακόμα, άλλα είδη πληροφορίας μπορούν να ληφθούν από ένα μέτρο οργάνωσης.

Στην εργασία [12] δίνεται μια ενδιαφέρουσα οπτική σχετικά με κάποιες γενικές συνθήκες που ένα καλό μέτρο οργάνωσης θα πρέπει να τηρεί:

1. Μοναδικότητα: Όταν και μόνο όταν εφαρμόζεται σε έναν χάρτη που διατηρεί τέλεια την τοπολογία, το μέτρο πρέπει να λαμβάνει μια ορισμένη μοναδική, ενδεικτική τιμή.

2. Συστηματική διάταξη: Όταν εφαρμόζεται σε μια σειρά χαρτών με μια συστηματικά μεταβαλλόμενη τοπολογία χαρτών, όπως μια μονοτονική αύξηση της διάστασης με σταθερό

όμως συνολικό αριθμό νευρώνων, η τιμή του μέτρου πρέπει επίσης να αλλάξει συστηματικά. Ως συνεπεία αυτής της συμπεριφοράς, το μέτρο πρέπει να εκφράσει ποιος από τους δύο χάρτες που δεν διατηρούν την τοπολογία, αντιπροσωπεύει καλύτερα τα δεδομένα.

3. Δυνατότητα αναπαραγωγής των αποτελεσμάτων: Για χάρτες όμοιους με εκείνους που παρήχθησαν από διαφορετικές αρχικοποιήσεις όμοια παραμετροποιημένων αλγορίθμων, το μέτρο πρέπει να παρουσιάζει ποσοτικά όμοια αποτελέσματα.

4. Σταθερότητα: Το μέτρο τοπολογίας πρέπει να εξαρτάται μόνο από την τοπολογία του χάρτη και όχι από κανονικοποίηση ή από μετασχηματισμούς συμμετρίας.

5.3 Μέτρα Οργάνωσης

Στην ενότητα αυτή παρουσιάζονται διαδοχικά τα μέτρα οργάνωσης και ποιότητας που χρησιμοποιήθηκαν στην εργασία μας. Η ανάλυσή που πραγματοποιείται, αφορά τόσο την αρχική μορφή τους όσο και αυτή που τελικά απέκτησαν αφού προσαρμόστηκαν κατάλληλα στη δομή και τις ανάγκες του υβριδικού SOHMMM.

5.3.1 Εντροπία

Στην εργασία [13] μελετήθηκε η μετάδοση πληροφορίας από την είσοδο στην έξοδο για μια ειδική κατηγορία νευρωνικών δικτύων πρόσθιας τροφοδότησης. Τα αποτελέσματα έδειξαν ότι υπό συγκεκριμένες συνθήκες η διαδικασία εκμάθησης αυτών των δικτύων μπορεί να παρατηρηθεί να εκφράζει μια μεγιστοποίηση της διατήρησης της πληροφορίας κατά τη μετάδοση δεδομένων από τους νευρώνες εισόδου στους νευρώνες εξόδου. Η τελευταία παρατήρηση συστήνει τη χρήση μιας βασισμένης στην θεωρία της πληροφορίας προσέγγισης για τη σχεδίαση των δικτύων, το λεγόμενο Linsker's informax principle με τις παραλλαγές του [14]. Οι αρχιτεκτονικές νευρωνικών δικτύων δεν έχουν σχεδιαστεί υποχρεωτικά ώστε να υπακούν πλήρως το informax principle, όμως παρατηρώντας τις ικανότητες τους στη διατήρηση της πληροφορίας μπορούμε πιθανώς να κατανοήσουμε καλύτερα τις ιδιότητες επεξεργασίας της πληροφορίας που διαθέτουν.

Η διαδικασία ενεργοποίησης ενός SOHMMM μπορεί να θεωρηθεί ως ένας μηχανισμός μετάδοσης πληροφορίας που μεταφέρει σήματα του V σε σήματα του A. Για αυτή τη διαδικασία μετάδοσης πληροφορίας μπορούμε να υπολογίσουμε πολλά είδη εντροπίας. Στην εργασία μας ασχοληθήκαμε με την εντροπία ενεργοποίησης νευρώνων ή απλώς εντροπία νευρώνων (neuron activation entropy).

Η εντροπία νευρώνων αποκτά μέγιστη τιμή όταν η πιθανότητα ενεργοποίησης είναι ίση για όλους τους νευρώνες, επομένως μέγιστη τιμή δείχνει ότι ο κβαντισμός του συνόλου εισόδου V πραγματοποιείται με τον βέλτιστο τρόπο. Βεβαίως, μόνο ο κβαντισμός αυτός προσδιορίζεται όσο το δυνατόν καλύτερα και όχι η θέση του σήματος εισόδου. Η απλή αυτή εντροπία αγνοεί εντελώς τις γεωμετρικές ιδιότητες της τοποθέτησης του πλέγματος στο χώρο εισόδου και η γνώση που μπορεί να μας παρέχει είναι καθαρά από την σκοπιά της θεωρίας της πληροφορίας. Εφόσον οι κόμβοι του χάρτη έχουν την ίδια πιθανότητα να ενεργοποιηθούν δεν παίζει ρόλο αν τα δεδομένα που θα περιγράφουν (καθένας από αυτούς) θα είναι περισσότερο (οδηγώντας σε υψηλό σφάλμα κβαντισμού) ή λιγότερο ανόμοια.

Κλασικός ορισμός. Έστω μια κατανομή σημάτων εισόδου. Αν η πιθανότητα να ενεργοποιηθεί ο νευρώνας i από ένα σήμα εισόδου είναι p_i, η εντροπία ενεργοποίησης νευρώνων δίνεται από τον τύπο

$$-\sum_{i \in A} p_i \log p_i \tag{5.1}$$

Ορισμός εξειδικευμένος για το SOHMMM. Στην περίπτωση του SOHMMM, επειδή πλέον δεν μπορούμε να μιλήσουμε για κατανομή σημάτων εισόδου, ως p_i θέσαμε τη μέση πιθανότητα να ενεργοποιηθεί ο νευρώνας i από το σύνολο των σημάτων (δεδομένων) εισόδου. Επίσης, πλέον όταν μιλάμε για πιθανότητα εννοούμε την πιθανότητα (likelihood) να εμφανίσει το μοντέλο HMM του νευρώνα i την κάθε ακολουθία των δεδομένων εισόδου. Αθροίζοντας για όλες αυτές τις πιθανότητες, λαμβάνουμε τη συνολική πιθανότητα ενεργοποίησης του νευρώνα i.

Εντέλει, ο τύπος της εντροπίας που χρησιμοποιήθηκε στις προσομοιώσεις μας ήταν

$$-\sum_{i \in A} p_i \log p_i \quad \textit{όπου} \quad p_i = \frac{1}{N} \sum_{O \in V} p_{iO} \tag{5.2}$$

Η εντροπία μπορεί να ιδωθεί ως η ποσότητα πληροφορίας Shannon που εκφράζεται από την απεικόνιση. Στην περίπτωση αυτή, οι ενεργοποιήσεις των μεμονωμένων νευρώνων είναι στοιχειώδεις και, από αυτή την άποψη, οι μεμονωμένες ενεργοποιήσεις δεν μπορούν να συγκριθούν η μία με την άλλη εφόσον δεν συμπεραίνουμε κάποια σημασιολογική δομή στο σύνολο των γεγονότων που έχουν μεταδοθεί και άρα και κάποια ομοιότητα. Πιο συγκεκριμένα, θα μπορούσαμε να πούμε πως η εντροπία δεν πρέπει να σχετίζεται καθόλου με την διατήρηση της τοπολογίας.

5.3.2 Σφάλμα κβαντισμού (quantization error)

Το νευρωνικό δίκτυο SOHMMM ανήκει, εκτός των άλλων, στην κατηγορία των αλγορίθμων που διαμερίζουν το σύνολο δεδομένων εισόδου (δεν θα ήταν ακριβές να μιλήσουμε για κβαντισμό διανυσμάτων, αφού τα δεδομένα εισόδου δεν είναι διανύσματα,αν και ο μηχανισμός είναι παρόμοιος) εφόσον μέσω αυτού κάθε στοιχείο του συνόλου των δεδομένων εισόδου (πχ ακολουθίες αλφάβητου,κλπ) απεικονίζεται (ή αντιστοιχίζεται) σε έναν από τους κόμβους ενός πλέγματος χαμηλής διάστασης. Συνεπώς, μπορούμε να ορίσουμε το σφάλμα κβαντισμού και για την περίπτωσή του. Το σφάλμα κβαντισμού στον κλασικό ορισμό του για τον κβαντισμό διανυσμάτων (vector quantization) αποτελεί το άθροισμα των αποστάσεων μεταξύ κάθε διανύσματος του χώρου εισόδου xi και του διανύσματος αναφοράς m_{xi} στο οποίο αντιστοιχίζεται, ήτοι

$$q_e = \frac{1}{N} \sum_{all\, x_i} [\![x_i - m_{xi}]\!]$$,όπου N ο αριθμός των δειγμάτων του συνόλου δεδομένων εισόδου

Στην περίπτωση του SOHMMM, η απόσταση μεταξύ δεδομένων εισόδου και αντίστοιχων κόμβων αναφοράς δεν μπορεί να οριστεί με τον κλασικό ορισμό. Αντί αυτού, ως μέτρο "απόστασης" μεταξύ των ακολουθιών παρατήρησης και των HMMs χρησιμοποιείται η

loglikelihood πιθανότητα (άλλωστε αυτό συμβαίνει και στον αλγόριθμο του SOHMMM προκειμένου να προσδιοριστεί η BMU - best matching unit). Κατ' αυτόν τον τρόπο, το σφάλμα κβαντισμού για το SOHMMM ορίζεται ως

$$q_e = \frac{1}{N \cdot M} \sum_{all\,O} -m_O.loglikelihood(O) \qquad (5.3)$$

Όπου με M συμβολίζουμε τον αριθμό των κόμβων του χάρτη και με N των αριθμό των δειγμάτων ακολουθιών παρατήρησης. Για τον υπολογισμό του σφάλματος κβαντισμού, απαιτείται να εισάγουμε μια ακόμη φορά στο δίκτυό μας το σύνολο των δειγμάτων δεδομένων εισόδου προκειμένου να προσδιοριστεί η απόσταση κάθε δείγματος από τον κόμβο (HMM) αναφοράς.

5.3.3 Τοπογραφικό σφάλμα (topographic error) [15]

Το τοπογραφικό σφάλμα είναι ένα μέτρο που έχει χρησιμοποιηθεί ευρέως στους αυτο-οργανούμενους χάρτες (SOMs) προκειμένου να ποσοτικοποιήσει την διατήρηση της τοπολογικής συνέχειας στον χάρτη. Στην περίπτωση αυτού του μέτρου, η έμφαση δίνεται στην αναλογία των δειγμάτων του συνόλου δεδομένων εισόδου που δείχνουν μια τοπική ασυνέχεια της χαρτογράφησης (απεικόνισης).

Ορισμός για το SOHMMM. Δοθείσης μιας ακολουθίας O που ανήκει στο χώρο εισόδου, έστω λ_i το πιο όμοιο και λ_j το δεύτερο πιο όμοιο HMM. Εάν οι αντίστοιχοι νευρώνες είναι γειτονικοί, τότε η απεικόνιση είναι τοπικά συνεχής, διαφορετικά υφίσταται μια τοπική ασυνέχεια, ή ένα τοπικό τοπογραφικό σφάλμα. Το τοπογραφικό σφάλμα για ολόκληρη την απεικόνιση λαμβάνεται αθροίζοντας τον αριθμό των τοπικών τοπογραφικών σφαλμάτων για όλα τα δείγματα και κανονικοποιώντας:

$$t_e = \frac{1}{N} \sum_{k=1}^{N} u(O_\kappa) \quad \text{με} \quad u(O_k) = \begin{cases} 1, & \text{μη γειτονικοι } i,j \\ 0, & \text{διαφορετικά} \end{cases} \qquad (5.4)$$

Μέσω αυτού του ορισμού, το τοπογραφικό σφάλμα μας δίνει μόνο μια ιδέα του ποιο τμήμα των τοπικών γειτονιών απεικονίζονται σωστά, χωρίς όμως να περιγράφει το είδος των εσφαλμένων απεικονίσεων. Για παράδειγμα, δοθέντων δύο σημείων πολύ κοντινών (όμοιων) στο χώρο εισόδου, δεν υφίσταται διαφορά για το τοπογραφικό σφάλμα μεταξύ του να απεικονιστούν με απόσταση ενός νευρώνα ή στις αντίθετες γωνίες του πλέγματος. Αυτό φαίνεται δικαιολογημένο αφού διαφορετικά θα ήταν δύσκολο να γνωρίζουμε αν μια υψηλή τιμή τοπογραφικού σφάλματος δείχνει ότι σχεδόν όλες οι γειτονιές περιέχουν τοπογραφικά σφάλματα μικρής ακτίνας ή ότι σε ορισμένα (λίγα) σημεία υπάρχει μια μεγάλη ασυνέχεια που απλώνεται σε όλο τον χάρτη

5.3.4 C-measure [3]

Έστω V_{in} ένας χώρος εισόδου και V_{out} ένας χώρος εξόδου. Έστω ότι με M συμβολίζουμε την αντιστοίχηση σημείων του V_{in} σε σημεία του V_{out} . Χρησιμοποιούμε τη λέξη «χώρος» υπό μια γενική έννοια εφόσον ο ένας ή και οι δύο από τους V_{in} και V_{out} μπορεί να μην έχουν κάποια γεωμετρική ερμηνεία. Έστω ότι για κάθε ένα από τους δύο χώρους υπάρχει μια συμμετρική συνάρτηση "ομοιότητας", η οποία για κάθε δοσμένο ζεύγος σημείων στον χώρο, προδιαγράφει με μια μη αρνητική αδιάστατη τιμή πόσο όμοια (ή ανόμοια) είναι αυτά τα σημεία. Έστω F η συνάρτηση για τον V_{in} και G για τον V_{out}. Ορίζουμε την συνάρτηση κόστους C ως εξής [3]:

$$C = \sum_{i=1}^{N} \sum_{j<i} F(i,j) G(M(i), M(j)) \tag{5.5}$$

όπου i και j είναι σημεία στο V_{in} και $M(i)$ και $M(j)$ είναι οι αντίστοιχες απεικονίσεις στον V_{out}. Το άθροισμα λαμβάνεται επί όλων των δυνατών ζευγών σημείων στο V_{in}. Εν γένει, ο ορισμός των F και G εξαρτάται από το συγκεκριμένο πρόβλημα. Μπορούν να είναι ευκλείδειες αποστάσεις σε κάποιο γεωμετρικό χώρο, κάποια (ενδεχομένως μη μονοτονική) συνάρτηση αυτών των αποστάσεων, ή κάποιο άλλο δοσμένο μέγεθος για το οποίο δεν είναι δυνατόν να θεωρηθεί ότι τα σημεία βρίσκονται σε κάποιο γεωμετρικό χώρο. Μια καλή απεικόνιση θεωρείται μια με υψηλή τιμή του C. Εντούτοις, αν μία από τις F ή G δίνεται ως συνάρτηση ανομοιότητας (δηλ. αυξάνεται με τη μείωση της ομοιότητας) τότε για μια καλή χαρτογράφηση το C έχει χαμηλή τιμή.

Επίσης, στην περίπτωση του SOHMMM οι συναρτήσεις F και G αναφέρονται στην απόσταση των HMM (μέσω του probabilistic distance measure) και στην απόσταση των νευρώνων στο πλέγμα αντίστοιχα. Αυτο συνιστά μια αρκετά ευρεία διαφοροποίηση στην αρχική φιλοσοφία του C, όπου οι αποστάσεις στον χώρο εισόδου λαμβανόταν μεταξύ των δεδομένων εισόδου, κάτι που δεν ήταν δυνατόν στην περίπτωσή μας αφού δεν έχουμε ορίσει ένα μέτρο απόστασης μεταξύ των ακολουθιών O. Η διαφοροποίηση αυτή πηγάζει από το γεγονός ότι πλέον οι κόμβοι αναφοράς και τα δεδομένα εισόδου αποτελούν δύο εντελώς διαφορετικά μοντέλα (HMMs και ακολουθίες αλφαβήτου αντίστοιχα). Στις προσομοιώσεις μας χρησιμοποιήσαμε ως F και G τις:

$$F(i,j) = D_s(\lambda_i, \lambda_j) \tag{5.6α}$$

$$G(M(i), M(j)) = |\vec{p}_{M(i)} - \vec{p}_{M(j)}| \tag{5.6β}$$

όπου με $p_{M(i)}$, $p_{M(j)}$ συμβολίζουμε τις θέσεις στο πλέγμα των νευρώνων i, j αντίστοιχα και Ds είναι το πιθανοτικό μέτρο απόστασης της εξίσωσης (3.42). Συνεπώς, τόσο η F όσο και η G είναι συναρτήσεις ανομοιότητας, πράγμα που σημαίνει πως η τιμή του C αυξάνεται όσο βελτιώνεται η ποιότητα της χαρτογράφησης.

Όπως είναι φανερό, το C μετρά τη συσχέτιση μεταξύ των F και G, συνεπώς θα μπορούσαμε να πούμε ότι το μέτρο C απαιτεί τον δεύτερο περιορισμό όπως περιγράφηκε στην εισαγωγή της 5.2, δηλαδή οι τιμές (αν)ομοιότητας μεταξύ των ζευγών των σημείων στον χώρο εισόδου και στον χώρο εξόδου πρέπει να συσχετίζονται τέλεια ενώ είναι επίσης απλό για να δειχθεί ότι εάν υπάρχει μια χαρτογράφηση που διατηρεί τη διάταξη, τότε μεγιστοποιώντας το C θα την βρούμε. Αυτό είναι ισοδύναμο με το ότι για δύο διανύσματα πραγματικών αριθμών, το εσωτερικό γινόμενό τους μεγιστοποιείται για όλες τις δυνατές αλλαγές μέσα στα δύο διανύσματα εάν τα στοιχεία των διανυσμάτων είναι όμοια

διατεταγμένα, μια απόδειξη του οποίου μπορεί να βρεθεί στο [16].

Σχέση του μεγέθους C με τα τετραγωνικά προβλήματα ανάθεσης (quadratic assignment problems). Η διατύπωση της διατήρησης της γειτονιάς από την άποψη του μέτρου *C* το θέτει μέσα στη κατηγορία τετραγωνικών προβλημάτων ανάθεσης (QAPs). Αυτά εμφανίζονται σε πολλά διαφορετικά πλαίσια, και λαμβάνουν τη μορφή εύρεσης της ελάχιστης ή μέγιστης τιμής μιας εξίσωσης παρόμοιας με αυτής του *C*. Ένα επεξηγηματικό παράδειγμα είναι η βέλτιστη σχεδίαση των πληκτρολογίων γραφομηχανών [17]. Εάν *F(i, j)* είναι ο μέσος χρόνος που χρειάζεται μια δακτυλογράφος για να πιέσει διαδοχικά τις θέσεις i και j στο πληκτρολόγιο, ενώ *G(p, q)* είναι η μέση συχνότητα με την οποία τα γράμματα p και q εμφανίζονται διαδοχικά στο κείμενο μιας δεδομένης γλώσσας (σημειώστε ότι σε αυτό το παράδειγμα οι *F* και *G* δεν είναι απαραιτήτως συμμετρικές), κατόπιν το πληκτρολόγιο που ελαχιστοποιεί το μέσο χρόνο δακτυλογράφησης θα είναι αυτό που ελαχιστοποιεί το γινόμενο

$$C = \sum_{i=1}^{N} \sum_{j=1}^{N} F(i,j)\, G(M(i), M(j))$$

με *M(i)* να συμβολίζει το γράμμα που βρίσκεται στη θέση i. Η θεωρία των QAP είναι άμεσα εφαρμόσιμη στο μέτρο *C*. Για παράδειγμα, η θεωρία των QAP παρέχει αρκετούς διαφορετικούς τρόπους υπολογισμού των ορίων των ελάχιστων και μέγιστων τιμών του *C* για κάθε πρόβλημα. Αυτό θα μπορούσε να είναι πολύ χρήσιμο για το πρόβλημα του προσδιορισμού της ποιότητας ενός χάρτη σχετικά με τον (άγνωστο) καλύτερο δυνατό (παρά απλά να κάνουμε μια σύγκριση μεταξύ δύο χαρτών). Μια ειδική περίπτωση είναι το όριο ιδιοτιμών [19]. Εάν οι ιδιοτιμές των συμμετρικών μητρών *F* και *G* είναι λi και μi αντίστοιχα, έτσι ώστε $\lambda 1 \leq \lambda 2 \leq ... \leq \lambda \nu$ και $\mu 1 \geq \mu 2 \geq ... \geq \mu \nu$ τότε μπορεί να δειχθεί ότι το άθροισμα $\sum(\lambda i\ \mu i)$ δίνει ένα κατώτατο όριο της τιμής του *C*.

Τα QAP είναι γενικά πολυπλοκότητας NP. Ένας μεγάλος αριθμός αλγορίθμων για ακριβείς και ευριστικές λύσεις έχει μελετηθεί [17], [19]. Εντούτοις, ιδιαίτερες αρχικοποιήσεις των F και G μπορούν να επιτύχουν πολύ αποδοτικούς αλγορίθμους για την εύρεση καλών τοπικών βέλτιστων. Τέτοιες εκτιμήσεις παρέχουν έναν πρόσθετο πρακτικό περιορισμό στο ποιες επιλογές *F* και *G* είναι οι περισσότερο κατάλληλες.

5.3.5 Μέτρο των Demartines - Herault [20]

Στην εργασία [20] προτείνεται ένα μέτρο οργάνωσης που επιχειρεί να ταιριάξει (ταυτίσει) τις ομοιότητες στον χώρο εισόδου και στον χώρο εξόδου ακριβώς. Η προσέγγιση αυτή δεν είναι καινούρια, αλλά έχει εισαχθεί αρχικά στις απεικονίσεις Sammon [21]. Η τελευταία μέθοδος, σε αντίθεση με το μέτρο Demartines-Herault, χρησιμοποιεί έναν όρο κανονικοποίησης και γι' αυτό είναι μη συμμετρική. Στην ανάλυσή μας, προτιμήθηκε το μέτρο των Demartines και Herault αφού, σύμφωνα με τους ίδιους, αποτελεί ένα πιο αποτελεσματικό αλγόριθμο ελαχιστοποίησης και παράγει καλύτερες αναπαραστάσεις "έντονα διπλωμένων δομών" σε σχέση με τις απεικονίσεις του Sammon.

Το μέτρο των Demartines και Herault ορίζεται ως

$$E = 1/2 \sum_{i=1}^{N} \sum_{j<i} (F(i,j) - G(M(i), M(j)))^2 h(G) \qquad (5.7)$$

όπου, όπως και προηγουμένως με i, j συμβολίζονται δυο τυχαίοι νευρώνες, με *M(i)* και *M(j)* οι απεικονίσεις τους στο πλέγμα και με *F* συμβολίζουμε μια συνάρτηση απόστασης στον χώρο εισόδου *V* και με *G* μια συνάρτηση απόστασης στον χώρο εξόδου *A*. (Σημειώνεται σε αυτό το σημείο πως οι συμβολισμοί στην αρχική δημοσίευση είναι αντίστοιχα *X* και *Y* για τις συναρτήσεις απόστασης απλώς προτιμήθηκε ένας ενιαίος συμβολισμός για απλότητα). Στην αρχική εργασία επιλέγονται η *F* να είναι η ευκλείδεια απόσταση και η *G* ομοίως. Η συνάρτηση *h* είναι μια μονοτονικά φθίνουσα συνάρτηση της απόστασης *G*, που χρησιμοποιείται προκειμένου να ενισχυθεί το ταίριασμα των αποστάσεων μεταξύ κοντινών νευρώνων στον χώρο εξόδου. Με αυτόν τον τρόπο, αν οι απεικονίσεις είναι αρκετά ανόμοιες ("μακριά" στο πλέγμα) δεν απασχολεί το μέτρο εάν τα i και j είναι όμοια. Στην αρχική εργασία η *h* επιλέγεται να είναι η συνάρτηση Gauss ή η βηματική συνάρτηση.

Για την περίπτωση του SOHMMM, χρησιμοποιήθηκαν οι συναρτήσεις (5.6α) και (5.6β) ενώ ως συνάρτηση *h* επιλέχτηκε η γκαουσιανή

$$h(G) = \exp(-G/\sigma^2) \tag{5.8}$$

με εύρος *σ* τέτοιο ώστε για κάθε τάξη γείτονα η *h* να μειώνεται κατά μια τάξη μεγέθους, δηλαδή για τους άμεσους γείτονες του κόμβου ισχύει *h=0.1*, για τους αμέσως επόμενους *h=0.01* και ούτω καθεξής.

5.3.6 Μέτρο παραμόρφωσης (distortion measure) [22] [23]

Το μέτρο παραμόρφωσης έχει χρησιμοποιηθεί ευρέως στην βιβλιογραφία των μέτρων οργάνωσης για τους αυτο-οργανούμενους χάρτες. Δύο λόγοι κυρίως το καθιστούν δημοφιλές: (1) στην περίπτωση διακριτού συνόλου δεδομένων και με σταθερό εύρος γειτονιάς, μπορεί να αποδειχθεί ότι είναι μια τοπική συνάρτηση ενέργειας του SOM [24], και (2) ο κανόνας εκπαίδευσης του SOM έχει αποδειχθεί ότι είναι μια υπολογιστικά αποδοτική προσέγγιση για την κλίση του μέτρου παραμόρφωσης [25]. Για τον αυτο-οργανούμενο χάρτη το μέτρο παραμόρφωσης ορίζεται ως:

$$E = \sum_{i=1}^{P} \sum_{j=1}^{N} h_{b_i j} [\![x_i - m_j]\!]^2 \tag{5.9}$$

όπου *P* ο αριθμός προτύπων, *N* ο αριθμός νευρώνων και *b* ο κόμβος νικητής για το διάνυσμα *xi* (best matching unit - BMU)

$$h_{b_i j} = \exp - [\![r_{bi} - r_j]\!]^2 / 2\sigma^2 \tag{5.10}$$

Στην περίπτωση του SOHMMM το μέτρο παραμόρφωσης μπορεί να οριστεί ως:

$$E = \sum_{i=1}^{P} \sum_{j=1}^{N} h_{b_i j} [\![m_j . loglikelihood(O_i)]\!]^2 \tag{5.11}$$

με όμοιους συμβολισμούς και συνάρτηση *h* αυτή της εξίσωσης (5.10). Η επιλογή της τιμής του *σ* είναι ίδια με αυτή του μέτρου των Demartines και Herault.

5.3.7 Τοπογραφικό γινόμενο (topographic product) [26]

Το τοπογραφικό γινόμενο προέρχεται από το χώρο της μη-γραμμικής δυναμικής και της ανάλυσης χρονικών σειρών [27]. Στη μελέτη των SOM εισήχθη από τους Bauer και Pawelzik [26] και είναι ένα από τα πρώτα μέτρα που προτάθηκαν προκειμένου να ποσοτικοποιήσουν τις ιδιότητες διατήρησης τοπολογίας του SOM. Το μέτρο αυτό περιγράφει την διατήρηση ή την παραβίαση των σχέσεων της γειτονιάς μεταξύ χαρτών πιθανώς διαφορετικής διάστασης. Παρουσιάζει ευαισθησία σε μεγάλης κλίμακας παραβιάσεις της διάταξης της γειτονιάς αλλά δεν προσμετρά τις παραμορφώσεις γειτονιάς που προκαλούνται από διαφορετικούς συντελεστές μεγέθυνσης. Στη μελέτη μας χρησιμοποιήθηκε προκειμένου να περιγράψει την ποιότητα της οργάνωσης για πλέγματα διαφορετικού μεγέθους. Στη αρχική του μορφή, αποκτά μηδενική τιμή όταν υπάρχει τέλεια διατήρηση τοπολογίας και αρνητική (θετική) τιμή όταν η διάσταση του χώρου εξόδου είναι πολύ μικρή (πολύ μεγάλη). Για τον υπολογισμό του, λαμβάνονται υπόψιν μόνο τα διανύσματα αναφοράς του χάρτη. Στην συνέχεια, παρουσιάζεται στην μορφή που λαμβάνει για την περίπτωση του SOHMMM.

Έστω ότι με $n_k^A(j)$ συμβολίζουμε τον k-οστό κοντινότερο γείτονα του κόμβου j, με τις αποστάσεις να μετρούνται στον χώρο εξόδου δηλαδή,

$$\mathbf{n_1^A(j):\ d^A(j,n_1^A(j)) = \min_{j'\in A-\{j\}} d^A(j,j')}$$

$$\mathbf{n_2^A(j):\ d^A(j,n_2^A(j)) = \min_{j'\in A-\{j,n_1^A(j)\}} d^A(j,j')}$$

$$\mathbf{\ldots}$$

Ομοίως, έστω ότι με $n_k^V(j)$ συμβολίζουμε τον k-οστό κοντινότερο γείτονα του κόμβου j, με τις αποστάσεις όμως να μετρούνται στον χώρο εισόδου δηλαδή,

$$\mathbf{n_1^V(j):\ d^V(\lambda_j,\lambda_{n_1^V(j)}) = \min_{j'\in A-\{j\}} d^V(\lambda_j,\lambda_{j'})}$$

$$\mathbf{n_2^V(j):\ d^V(\lambda_j,\lambda_{n_1^V(j)}) = \min_{j'\in A-\{j,n_1^V(j)\}} d^V(\lambda_j,\lambda_{j'})}$$

$$\mathbf{\ldots}$$

Σε μια άλλη διατύπωση, μπορούμε να πούμε ότι ουσιαστικά ορίζουμε μια διάταξη στον χώρο εισόδου και στον χώρο εξόδου για κάθε κόμβο j:

$$\boldsymbol{d_A(j,n_1^A(j)) \le d_A(j,n_2^A(j)) \le \ldots \le d_A(j,n_{N-1}^A(j))} \qquad \textbf{(5.12α)}$$

και

$$\boldsymbol{d_V(\lambda_j,\lambda_{n_1^V(j)}) \le d_V(\lambda_j,\lambda_{n_2^V(j)}) \ldots \le d_V(\lambda_j,\lambda_{n_{N-1}^V(j)})} \qquad \textbf{(5.12β)}$$

Ως μέτρο απόστασης στο χώρο εξόδου λαμβάνουμε την ευκλείδεια απόσταση και στο χώρο εισόδου το πιθανοτικό μέτρο απόστασης μεταξύ HMM. Με βάση τους ανωτέρω

συμβολισμούς, ορίζουμε στη συνέχεια τους λόγους

$$Q_1(j,k) = \frac{\left|d_V(\lambda_j, \lambda_{n_k^A(j)})\right|}{\left|d_V(\lambda_j, \lambda_{n_k^V(j)})\right|} \quad \textbf{(5.13α)}$$

$$Q_2(j,k) = \frac{d_A(j, n_k^A(j))}{d_A(j, n_k^V(j))} \quad \textbf{(5.13β)}$$

Με βάση τους ορισμούς αυτούς, θα ισχύει $Q_1(j,k)=Q_2(j,k)=1$ μόνο εάν οι κοντινότεροι γείτονες τάξεως k στον χώρο εισόδου και στον χώρο εξόδου ταυτίζονται. Σε ό,τι αφορά το Q_1 είναι πιθανό να υπάρχει μια πολύ μικρή απόκλιση από τη μονάδα, αφού το dv, λόγω του ορισμού του (βλέπε κεφάλαιο 3) είναι πιθανό για τα ίδια ΗΜΜ να εμφανίσει μια αμελητέα διαφορά στο μέτρο της απόστασής τους μεταξύ δύο υπολογισμών (ανάλογα με το μέγεθος των παραγόμενων ακολουθιών O). Σημειώνουμε εδώ, πως γι' αυτό το λόγο, ιδιαίτερα για μεγάλης διάστασης πλέγματα, πρέπει να δοθεί ιδιαίτερη προσοχή στο μήκος των παραγόμενων ακολουθιών για το πιθανοτικό μέτρο απόστασης. Εν γένει πάντως, κάθε απόκλιση των Q_1 και Q_2 από το 1 δείχνει μια παραβίαση της διάταξης των πιο κοντινών γειτόνων από τον χάρτη. Εντούτοις, αυτό αποτελεί ένα πολύ ευαίσθητο μέτρο για τη διατήρηση των σχέσεων της γειτονιάς, αφού ένας τοπικά τεντωμένος (ή παραμορφωμένος) χάρτης μπορεί να διατηρήσει της σχέσεις γειτονιάς ακόμα και αν η διάταξη των πιο κοντινών γειτόνων παραβιάζεται. Το πρόβλημα αυτής της ευαισθησίας μπορεί να ξεπεραστεί πολλαπλασιάζοντας τα μεγέθη Q για όλες τις τάξεις k. Με κατάλληλη κανονικοποίηση λαμβάνουμε

$$P_1(j,k) = \left(\prod_{l=1}^{k} Q_1(j,l)\right)^{\frac{1}{k}} \quad \textbf{(5.14α)}$$

$$P_2(j,k) = \left(\prod_{l=1}^{k} Q_2(j,l)\right)^{\frac{1}{k}} \quad \textbf{(5.14β)}$$

Για τις νέες μεταβλητές P_1 και P_2 ισχύει

$$P_1(j,k) \geqslant 1$$

$$P_1(j,k) \leqslant 1$$

Στους όρους P_1 και P_2 ακυρώνεται μια διαφορετική διάταξη κοντινότερων νευρώνων, εφόσον οι πρώτοι k κοντινότεροι γείτονες του j στο V και στο A ταυτίζονται (ανεξάρτητα της διάταξής τους). Τα P_1 και P_2 είναι ευαίσθητα μόνο σε σοβαρές παραβιάσεις της γειτονιάς, για παράδειγμα εάν δυο κόμβοι βρίσκονται κοντά στον χώρο εισόδου αλλά οι αντίστοιχες θέσεις τους βρίσκονται μακριά στον χώρο εξόδου.

Σταθεροί συντελεστές μεγέθυνσης του χάρτη δεν μεταβάλλουν τις διατάξεις άμεσων γειτόνων, οπότε τα γινόμενα P_1 και P_2 έχουν την σημαντική ιδιότητα να μην επηρεάζονται από σταθερές κλίσεις του χάρτη. Κατανομές εισόδου με χωρικές διαφοροποιήσεις προκαλούν διαφοροποιήσεις των συντελεστών μεγέθυνσης του χάρτη, που αντιστοιχούν σε μη μηδενικές παραγώγους δεύτερης τάξης.

Μια μεταβολή στον τοπικό συντελεστή μεγέθυνσης μπορεί να προκαλέσει αλλαγές στις διατάξεις των άμεσων γειτόνων. Εφόσον αυτές οι συνεισφορές δεύτερης τάξης αποκτούν

έναν μέσο όρο τοπικά, τα γινόμενα P_1 και P_2 παραμένουν κοντά στη μονάδα το καθένα για επαρκώς μεγάλες τιμές του k. Για την αντίθετη περίπτωση, όπου οι δεύτερες παράγωγοι δεν αποκτούν έναν μέσο όρο τοπικά, συνδυάζουμε τα P_1 και P_2 πολλαπλασιάζοντάς τα και λαμβάνουμε

$$P_3(j,k) = \left(\prod_{l=1}^{k} Q_1(j,l)Q_2(j,l)\right)^{\frac{1}{2k}} \tag{5.18}$$

Το τελευταίο αυτό βήμα έχει ως αποτέλεσμα, λόγω της αντίστροφης συμπεριφοράς των P_1 και P_2, οι συνεισφορές των καμπυλοτήτων να καταστέλλονται ενώ οι παραβιάσεις των γειτονιών να ανιχνεύονται με $P_3 \neq 1$. Ένας πρόσθετος σημαντικός λόγος για τον ορισμό του P_3, εντούτοις είναι ότι $P_1 > 1/P_2$, όταν ο χώρος εισόδου συμπτύσσεται στον χώρο εξόδου και $P_1 < 1/P_2$, όταν ο χώρος εξόδου συμπτύσσεται στον χώρο εισόδου. Επομένως, τιμή μεγαλύτερη ή μικρότερη του 1 για το P_3 δείχνει ότι η διάσταση του χώρου εξόδου είναι πολύ μεγάλη ή πολύ μικρή αντίστοιχα.

Απομένει να οριστεί ένας κατάλληλος μέσος όρος για το $P_3(j, k)$. Για αυτόν τον σκοπό υπάρχουν αρκετές εναλλακτικές. Καταρχήν, θα μπορούσαμε να εξετάσουμε σημειακά όλους τους κόμβους j έτσι ώστε να αποκτήσουμε μια χωρική εκτίμηση των παραβιάσεων της τοπολογίας. Ένας δεύτερος τρόπος είναι να κατασκευάσουμε ιστογράμματα ώστε να λάβουμε την πιθανότητα ευρέων παραβιάσεων της γειτονιάς μαζί με μια εκτίμηση της ολικής διακύμανσης του P_3. Ο πιο απλός τρόπος λήψης ενός μέσου όρου, εντούτοις, συνίσταται στο να αθροίσουμε επί όλων των κόμβων και όλων των δυνατών τάξεων γειτονιάς, μια μέθοδος που αρκεί για τις περισσότερες πρακτικές εφαρμογές. Ενδιαφερόμενοι μόνο για την απόκλιση από τη μονάδα, λαμβάνουμε την μέση τιμή του P_3 και τελικά καταλήγουμε στον τύπο του P

$$P = \frac{1}{N(N-1)} \sum_{j=1}^{N} \sum_{k=1}^{N-1} \log(P_3(j,k)) \tag{5.19}$$

Ένα πλεονέκτημα του τοπογραφικού προϊόντος είναι ότι η τιμή του P μπορεί να χρησιμοποιηθεί για να συνοψίσει την ποιότητα της διατήρησης τοπολογίας με έναν μοναδικό αριθμό, ενώ οι πραγματικές αιτίες μπορούν να εξεταστούν από τις καμπύλες χάραξης για $P1$, $P2$ και $P3$, ή με την απεικόνιση των λαθών ανά μονάδα χάρτη σε ένα πλέγμα SOM. Το τοπογραφικό γινόμενο μπορεί να υπολογιστεί και για χάρτες που εκπαιδεύονται με σύνολα δεδομένων που περιέχουν ελλιπείς τιμές, αφού το διάνυσμα αναφοράς αυτό καθεαυτό δεν περιέχει ποτέ ελλιπείς τιμές.

Σχόλιο. Όπως έχει αναφερθεί και προηγουμένως, μια σημαντική ιδιότητα του μέτρου αυτού είναι ότι για τον υπολογισμό του απαιτούνται μόνο τα διανύσματα αναφοράς και δεν χρειάζεται να γνωρίζουμε τη μορφή των δεδομένων εισόδου. Επίσης, στο πρώτο βήμα, η επιλογή της διάταξης $n_i^A(j)$ δεν είναι μοναδική, αφού υπάρχουν διάφορα ζεύγη νευρώνων για τα οποία $\mathbf{d^A(j,k) = d^A(j,k')}$, $\mathbf{k \neq k'}$. Συνεπώς διαφορετικές επιλογές της διάταξης μπορεί να οδηγήσουν σε διαφορετικές τιμές για τις ενδιάμεσες ποσότητες Q_1 ,Q_2 ενώ, εμπειρικά, το τελικό αποτέλεσμα φαίνεται να είναι αρκετά ανεξάρτητο από αυτές τις επιλογές.

Σε αυτό το σημείο βέβαια, πρέπει να σημειωθεί ότι το τοπογραφικό γινόμενο, τουλάχιστον για την περίπτωση του SOHMMM, έχει πάρα πολύ μεγάλες υπολογιστικές απαιτήσεις. Συγκρινόμενο με τα άλλα μέτρα, ο χρόνος υπολογισμού του ήταν κατά αρκετές τάξεις μεγέθους ($>10^3$) μεγαλύτερος, κάτι που καθιστά συχνά σαφώς προτιμητέα τα άλλα

μέτρα. Εν γένει, (βλέπε και κεφάλαιο 6) θα μπορούσαμε να πούμε ότι είναι προτιμητέο μόνο στην περίπτωση διερεύνησης του βέλτιστου μεγέθους του πλέγματος, χωρίς αυτό όμως να σημαίνει ότι έχει και καλύτερη απόδοση σε σχέση με άλλα κριτήρια.

5.3.8 Το κριτήριο Kalinski - Harabasz [28]

Πέραν των μέτρων οργάνωσης, στην εργασία αυτή μελετήθηκαν και κριτήρια για την αναγνώριση συστάδων σε ευκλείδειους χώρους μεγάλων διαστάσεων. Η αρχική χρήση αυτών των δεικτών δεν σχετίζεται με τον αλγόριθμο του SOM αλλά με μεθόδους (όπως ο αλγόριθμος των κ-μέσων) στη λογική της εύρεσης του πλήθους των συστάδων ενός συνόλου δεδομένων και ταξινόμησής του εν λόγω συνόλου σε αυτές τις συστάδες. Η βιβλιογραφία των δεικτών έγκυρων συστάδων (cluster validity indexes) είναι αρκετά πλούσια και ιδιαίτερο ενδιαφέρον παρουσιάζει η εργασία [29] όπου επιχειρείται μια σύγκριση 30 τέτοιων κριτηρίων. Αρκετά από αυτά τα κριτήρια [30]-[34] δοκιμάστηκαν στην εργασία μας με μία χρήση διαφορετική της αρχικής τους, προκειμένου να ποσοτικοποιήσουν την διαδικασία της αυτο-οργάνωσης του SOHMMM. Αν και η συμπεριφορά αρκετών από αυτά παρουσίαζε μια μονοτονικότητα (η τιμή του κριτηρίου αυξανόταν ή μειωνόταν στον χρόνο) εντέλει η απόδοσή τους δεν ήταν σε καμία περίπτωση συγκρίσιμη με αυτή των, πλέον ειδικευμένων, μέτρων οργάνωσης. Στην ενότητα αυτή θα εισάγουμε απλώς ενδεικτικά το κριτήριο των Kalinski και Harabasz που από πολλούς ερευνητές έχει θεωρηθεί το πλέον επιτυχημένο, προκειμένου να δείξουμε ότι είναι δυνατή η χρήση αυτών των κριτηρίων (ίσως με κάποιες τροποποιήσεις) για έναν διαφορετικό σκοπό, αυτόν της ποσοτικοποίησης της οργάνωσης. Το έναυσμα για την δοκιμή αυτών των κριτηρίων δόθηκε από το γεγονός ότι στο SOHMMM κάθε κόμβος αποτελεί μια συστάδα, εφόσον τα δεδομένα εισόδου περιγράφονται το καθένα από έναν κόμβο αναφοράς.

Το κριτήριο αυτό ορίζεται ως

$$HC = \frac{BGSS}{k-1} \Big/ \frac{WGSS}{n-k} \tag{5.21}$$

όπου με *k* συμβολίζεται ο αριθμός των συστάδων(κόμβων), με n συμβολίζεται ο αριθμός των δεδομένων εισόδου, με *BGSS* συμβολίζεται το άθροισμα των αποστάσεων μεταξύ των συστάδων (between group sum of squares), με *WGSS* συμβολίζεται το άθροισμα των αποστάσεων εντός κάθε συστάδας (between group sum of squares), ήτοι μεταξύ του κόμβου και των ακολουθιών-δεδομένων που αυτός περιγράφει. Για την περίπτωση του SOHMMM, ορίζουμε συνεπώς

$$BGSS = \sum_{i=1}^{i=k} \sum_{j<i} D(\lambda_i, \lambda_j) \tag{5.22}$$

$$WGSS = \sum_{i=1}^{i=k} \sum_{j=1}^{j=n_k} \lambda_i . loglikelihood(O_j) \tag{5.23}$$

5.3.9 Δείγματα εισόδου που αλλάζουν κόμβο αναφοράς μεταξύ δύο επαναλήψεων

Ένα τελευταίο, πολύ απλό μέτρο που εισάγαμε προκειμένου να μπορούμε να παρατηρήσουμε την διαδικασία της αυτο-οργάνωσης είναι το άθροισμα των ακολουθιών του συνόλου εισόδου που αλλάζουν κόμβο αναφοράς μεταξύ δύο διαδοχικών εποχών. Το μέτρο αυτό ορίζεται επομένως ως:

$$C_e = \frac{1}{N}\sum_{k=1}^{N} p(O_k) \quad \text{με} \quad p(O_k^e) = \begin{cases} 1, & BMU(O_k^e) \neq BMU(O_k^{e-1}) \\ 0, & \textit{διαφορετικά} \end{cases} \tag{5.3}$$

Όπου με *e* συμβολίζουμε την επανάληψή (εποχή) *e* και με BMU τον κόμβο νικητή για τη δοσμένη ακολουθία (best matching unit).Το μέτρο αυτό είναι σε θέση να δείξει με επιτυχία την πορεία της εκπαίδευσης του SOHMMM αφού αναμένεται με την πάροδο των εποχών και την ολοένα και καλύτερη διάταξη του χάρτη ο αριθμός των ακολουθιών που αλλάζουν ετικέτα να μειώνεται.

5.4 Μέτρα οργάνωσης που μελετήθηκαν αλλά δεν χρησιμοποιήθηκαν

Στην περίπτωση του SOHMMM ένα πολύ σημαντικό ζήτημα αποτελεί το γεγονός ότι δεν υπάρχει κάποιο γενικό μέτρο απόστασης μεταξύ των δεδομένων εισόδου. Ένα τέτοιο μέτρο δεν έχει οριστεί για τα δεδομένα που χειρίζεται ένα HMM. Στην περίπτωση των ακολουθιών κάποιου αλφαβήτου υφίστανται κάποια μέτρα απόστασης [35], [36] όμως η λογική χρήσης τους απέχει αρκετά από την πιθανοτική αντιμετώπιση του SOHMMM. Επίσης, όπως έχει ειπωθεί προηγουμένως, τα δεδομένα εισόδου στο SOHMMM δεν υφίστανται σε κάποιο χώρο, συνεπώς δεν ήταν δυνατή η προσέγγιση του χώρου εισόδου με κελιά Voronoi και η χρήση γράφων Dealaunay και των συναφών με αυτές τις προσεγγίσεις μέτρων οργάνωσης.

Για τους παραπάνω λόγους δεν ήταν δυνατή η εφαρμογή της δημοφιλούς τοπογραφικής συνάρτησης των Vilmann et al. [2], του επίσης αρκετά γνωστού στη βιβλιογραφία των SOM μέτρου οργάνωσης των Kaski και Lagus [38] καθώς και των μέτρων εμπιστοσύνης (trustworthiness) και διατήρησης γειτονιάς (neighborhood preservation) των Kaski και Venna [37]. Τα μέτρα αυτά χρησιμοποιούν είτε την έννοια της διαίρεσης του χώρου εισόδου σε κελιά Voronoi [2], είτε την έννοια της διάταξης των δεδομένων του χώρου εισόδου [38], είτε προϋποθέτουν την ύπαρξη τριγωνικής ανισότητας στον χώρο εισόδου [37], έννοιες που δεν είναι δυνατόν να εφαρμοστούν στο SOHMMM.

Κεφάλαιο 6
Προσομοιώσεις για τα μέτρα οργάνωσης του SOHMMM

6.1 Εισαγωγή

Στο κεφάλαιο αυτό παρουσιάζονται και αναλύονται τα αποτελέσματα των προσομοιώσεων που πραγματοποιήθηκαν ενώ επιχειρείται και μια σύγκρισή των ληφθέντων αποτελεσμάτων για τα διάφορα μέτρα οργάνωσης. Η σύγκριση αυτή γίνεται υπό το πρίσμα ότι τα διαφορετικά μέτρα είναι πλέον κατάλληλα, το καθένα, σε διαφορετικές εργασίες, πχ στην περιγραφή της μονοτονίας της εκπαίδευσης ή στην απόφανση σχετικά με την επιλογή βέλτιστου αριθμού καταστάσεων για τα μαρκοβιανά μοντέλα,κοκ. Χρησιμοποιήθηκαν τρία χαρακτηριστικά τεχνητά σύνολα δεδομένων και δύο πραγματικά (πρωτεϊνικές ακολουθίες). Οι προσομοιώσεις έγιναν τόσο από την άποψη της περιγραφής της διαδικασίας της εκπαίδευσης (μονοτονία) όσο και από την απόφανση σχετικά με τις βέλτιστες παραμετροποιήσεις του SOHMMM όσον αφορά το μέγεθος του πλέγματος, τον αριθμό των απαιτούμενων εποχών εκπαίδευσης και τον αριθμό καταστάσεων των HMMs. Η μελέτη αυτή, είχε συνολικά ως στόχο να καταδείξει ότι τα κριτήρια που αναπτύχθηκαν στο κεφάλαιο 5 είναι σε θέση να ποσοτικοποιήσουν με επιτυχία την οργάνωση ενός SOHMMM κάνοντας δυνατή έτσι την βέλτιστη παραμετροποίησή του, την εποπτεία της διαδικασίας εκπαίδευσης ή τον πρόωρο τερματισμό λειτουργίας του. Στη συνέχεια, παρουσιάζονται αρχικά τα σύνολα δεδομένων που χρησιμοποιήθηκαν και έπειτα τα αποτελέσματα των μετρήσεων για τις διαφορετικές επιλεχθείσες λειτουργίες για το κάθε ένα σύνολο δεδομένων.

6.2 Σύνολα δεδομένων

Στην ενότητα αυτή παρουσιάζονται αναλυτικά τα χαρακτηριστικά των συνόλων δεδομένων που χρησιμοποιήθηκαν. Στην μελέτη μας επιχειρήθηκε να χρησιμοποιηθούν μικρού μεγέθους σχετικά, πλην όμως χαρακτηριστικά σύνολα δεδομένων. Με αυτόν τον τρόπο κατέστη δυνατόν να πραγματοποιηθεί μια εκτενής ανάλυση των δυνατοτήτων και επιδόσεων αυτών των κριτηρίων για αρκετά χαρακτηριστικές περιπτώσεις δεδομένων, σε ένα εύλογο παράλληλα χρονικό διάστημα. Θα πρέπει σε αυτό το σημείο να τονιστεί ότι η επιλογή των συνόλων δεδομένων δεν ήταν μια εύκολη διαδικασία καθώς δεν υφίστανται κάποια benchmark datasets για το SOHMMM ή για συναφείς προσεγγίσεις αλγόριθμων εκμάθησης.

6.2.1 Τεχνητά σύνολα δεδομένων

Στις προσομοιώσεις μας χρησιμοποιήθηκαν τρία τεχνητά σύνολα δεδομένων που παρουσιάζονται αναλυτικά ακολούθως:

FirstArtificial(seqs).txt

Το πρώτο από τα τεχνητά σύνολα δεδομένων έχει μια σχετικά απλή δομή, προκύπτοντας από HMM μιας κατάστασης. Σκοπός είναι να υπάρχει μια ξεκάθαρη διάταξη που να επιτυγχάνεται σε λιγα βήματα εκπαίδευσης. Το αρχείο αποτελείται από ένα σύνολο από 150 συμβολοακολουθίες μεταβλητού μήκους (100-200), που παρήχθησαν από τρία διαφορετικά HMM μίας κατάστασης. Το αλφάβητο των τριών HMM είναι το {A, G, C, T}. Οι αντίστοιχοι πίνακες εκπομπής απεικονίζονται παρακάτω:

b1 =[53.0 23.0 12.0 12.0]
b_2 =[13.0 3.0 22.0 62.0]
b_3 =[4.0 69.0 3.0 24.0]

Από κάθε HMM ελήφθησαν 50 δείγματα, συνεπώς τα δεδομένα οργανώνονται σε τρεις ισομεγεθείς συστάδες.

SecondArtificial(seqs).txt

Στην περίπτωση του δεύτερου τεχνητού συνόλου δεδομένων δόθηκε έμφαση στην ύπαρξη δεδομένων που θα παράγονται από HMM περισσοτέρων καταστάσεων που θα έχουν το ίδιο αλφάβητο. Ένα σύνολο από 40 συμβολοακολουθίες μεταβλητού μήκους (200-500), παρήχθησαν από τρία HMM δέκα καταστάσεων. Το αλφάβητο των τριών HMM είναι το {M, N, B,V, C, X}. Θεωρώντας τρία υποσύνολα, τα *A1 = {MN}, A2 = {BV}, A3 = {CX}* υποθέτουμε και για τα τρία HMM μια τυχαία εναλλαγή μεταξύ των καταστάσεων ως εξής: σύμφωνα με μια τυχαία διαδικασία επιλέγεται ένα από τα υποσύνολα. Από αυτό επιλέγεται μια παρατήρηση σύμφωνα με την κατανομή πιθανότητας (πχ μπορεί το M να έχει πιθανότητα εμφάνισης 0.4 και το N 0.6 για το *A1*). Το επόμενο υποσύνολο επιλέγεται τυχαία σύμφωνα με μια γεννήτρια τυχαίων αριθμών. Όπως παρατηρούμε και παρακάτω, για κάθε ένα από τα HMM ένα υποσύνολο έχει σαφώς μεγαλύτερη πιθανότητα εμφάνισης (πχ για την περίπτωση του πρώτου HMM επιλέγεται με πιθανότητα 0.8 μια παρατήρηση από το σύνολο *A2* και με 0.1 από τα *A1, A3* αφού 8 από τις 10 καταστάσεις παράγουν αυτές τις ακολουθίες). Ακολουθούν οι πίνακες εκπομπής:

b_1 =[0.5 0.5 0.0 0.0 0.0 0.0
0.0 0.0 0.5 0.5 0.0 0.0
0.0 0.0 0.4 0.6 0.0 0.0
0.0 0.0 0.47 0.53 0.0 0.0
0.0 0.0 0.45 0.55 0.0 0.0
0.0 0.0 0.57 0.43 0.0 0.0
0.0 0.0 0.59 0.41 0.0 0.0
0.0 0.0 0.5 0.5 0.0 0.0
0.0 0.0 0.5 0.5 0.0 0.0
0.0 0.0 0.0 0.0 0.5 0.5]

b_2 =[0.0 0.0 0.0 0.0 0.55 0.45
0.0 0.0 0.0 0.0 0.5 0.5
0.0 0.0 0.0 0.0 0.5 0.5
0.0 0.0 0.0 0.0 0.55 0.45
0.0 0.0 0.0 0.0 0.5 0.5
0.0 0.0 0.0 0.0 0.5 0.5
0.0 0.0 0.0 0.0 0.47 0.53
0.0 0.0 0.0 0.0 0.5 0.5
0.5 0.5 0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0 0.42 0.58]

b_3 =[0.5 0.5 0.0 0.0 0.0 0.0
0.5 0.5 0.0 0.0 0.0 0.0
0.53 0.47 0.0 0.0 0.0 0.0
0.5 0.5 0.0 0.0 0.0 0.0

0.5	**0.5**	**0.0**	**0.0**	**0.0**	**0.0**
0.49	**0.51**	**0.0**	**0.0**	**0.0**	**0.0**
0.5	**0.5**	**0.0**	**0.0**	**0.0**	**0.0**
0.5	**0.5**	**0.0**	**0.0**	**0.0**	**0.0**
0.0	**0.0**	**0.0**	**0.0**	**0.5**	**0.5**
0.44	**0.56**	**0.0**	**0.0**	**0.0**	**0.0]**

Στην περίπτωση αυτή ελήφθησαν 10 δείγματα από κάθε ένα από τα δύο πρώτα datasets και 20 δείγματα από το τρίτο σύνολο δεδομένων.

ThirdArtificial(seqs).txt

Το τρίτο σύνολο δεδομένων αποτελείται από 30 δείγματα. Στην περίπτωση αυτή μας ενδιέφερε η διακριτική ικανότητα του SOHMMM όσον αφορά δεδομένα που παρήχθησαν από HMM διαφορετικού αριθμού καταστάσεων. Έτσι, τα δεδομένα παράγονται από ένα HMM τριών καταστάσεων (>XC->VB->NM-) και από δύο κατανομές (με αλφάβητα MNBV και JKLZ αντίστοιχα) για της οποίες η πιθανότητα εμφάνισης κάθε συμβόλου σε κάθε παρατήρηση είναι ισοπίθανη. Ουσιαστικά, οι δύο τελευταίες κατανομές δημιουργούνται από HMM μίας κατάστασης με ισοπίθανες εκπομπές για κάθε σύμβολο. Λαμβανονται 10 δείγματα από κάθε κατανομή (συνολικά 30) ενώ κάθε ακολουθία έχει μήκος μεταξύ 200 και 500.

6.2.2 Πραγματικά σύνολα δεδομένων

Τα πραγματικά σύνολα δεδομένων που χρησιμποποιήθηκαν, αφορούν πρωτεινικές ακολουθίες. Τα δεδομένα αυτά χαρακτηρίζονται από μεταβλητού, αρκετά μεγάλου μεγέθους ακολουθίες συμβόλων που επιπρόσθετα λαμβάνονται από ένα αλφάβητο 20 χαρακτήρων (αμινοξέα).

Globins

Τα globins είναι μια πολύ γνωστή οικογένεια πρωτεινών που περιέχουν αιμίνη και που συνδέονται με το οξυγόνο και εμπλέκονται στη μεταφορά και αποθήκευσή του. Η οικογένεια αυτή περιλαμβάνει πολλές υπο-οικογένειες. Από κρυσταλλογραφικές μελέτες, όλες οι globins έχουν παρόμοιες τρισδιάστατες δομές αλλά ευρέως αποκλίνουσες ακολουθίες. Οι ακολουθίες που χρησιμοποιήθηκαν εδώ εξήχθησαν από την πρωτεϊνική βάση γνώσεων iProClass [26], μια βάση δεδομένων που ενσωματώνει στοιχεία και πληροφορίες πάνω από 90 βιολογικών βάσεων δεδομένων. Συνολικά εξήχθησαν 560 πρωτεΐνες που ανήκουν στα globins, οργανωμένες σε τρεις υποκατηγορίες (συγκεκριμένα α-chain hemoglobins, b-chain hemoglobins και myoglobins). Η προκύπτουσα σύνθεση του συνόλου των δεδομένων είναι 194 α-globins, 216 β-globins, και 150 myoglobins.

Στο πρώτο σύνολο δεδομένων , χρησιμοποιήθηκαν συνολικά 30 πρωτεΐνες, 10 από κάθε κατηγορία. Σκοπός, ήταν η εξαγωγή συμπερασμάτων σχετικά με τις δυνατότητες των κριτηρίων οργάνωσης όσον αφορά πραγματικά δεδομένα, με χρήση μικρότερων διαστάσεων χαρτών. Στο δεύτερο σύνολο δεδομένων χρησιμοποιήθηκαν όλες οι πρωτεΐνες (560). Σε όλες τις περιπτώσεις τα δεδομένα οργανώθηκαν με πολύ ικανοποιητικό τρόπο (οι απεικονίσεις του

χάρτη ήταν επιτυχημένες) ενώ το σύνολο των κριτηρίων οργάνωσης είχε επίσης πολύ ικανοποιητική συμπεριφορά.

6.3 Προσομοιώσεις

Στην ενότητα αυτή παρουσιάζονται και αναλύονται τα αποτελέσματα των προσομοιώσεων που πραγματοποιήθηκαν. Μελετήθηκε αρχικά η συμπεριφορά κάθε κριτηρίου (πλην του τοπογραφικού γινόμενου που) κατά τη διάρκεια της εκπαίδευσης διερευνώντας την ύπαρξη κάποιας μονοτονίας σε αυτήν. Στη συνέχεια, μελετήθηκε η δυνατότητα των μέτρων οργάνωσης να αποφανθούν περί του βέλτιστου πλέγματος, αριθμού καταστάσεων και αριθμού εποχών εκπαίδευσης. Η μελέτη αυτή θα μπορούσε να επεκταθεί περιλαμβάνοντας και άλλες παραμέτρους του SOHMMM (αρχικό-τελικό learning rate, αρχικές καταστάσεις των HMM) όμως κύριος στόχος μας ήταν να αποδειχθεί η δυνατότητα των μέτρων οργάνωσης να ποσοτικοποιήσουν την οργάνωση του SOHMMM και όχι να εξετάσουμε κάθε δυνατή χρήση των μέτρων αυτών. Στο Παράρτημα Ε παρουσιάζονται τα αποτελέσματα κάθε προσομοίωσης, δηλαδή ποια πρότυπα περιέχει ο κάθε κόμβος του SOHMMM μετά το πέρας της εκπαίδευσης.

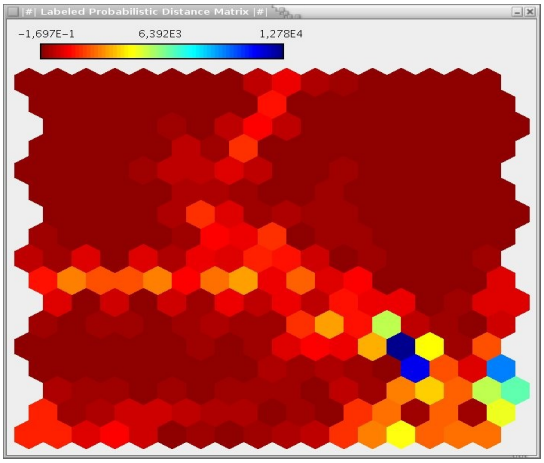
Σημείωση. Το τοπογραφικό γινόμενο χρησιμοποιήθηκε μόνο στην ποσοτικοποίηση της οργάνωσης για διαφορετικά μεγέθη πλέγματος, καθώς αποδείχθηκε πάρα πολύ απαιτητικό σε χρόνο. Εξάλλου, η αρχική χρήση του έγινε μόνο ως προς την επιλογή διάστασης πλέγματος, ενώ σε κάθε περίπτωση η επίδοση του ήταν σε αρκετές περιπτώσεις μη ικανοποιητική.

6.3.1 Μονοτονία εκπαίδευσης

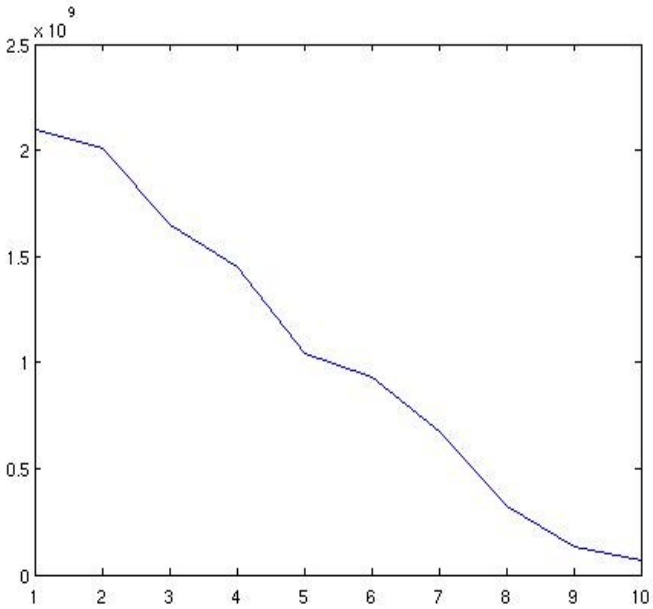
6.3.1.1 Τεχνητά σύνολα δεδομένων

FirstArtificial(seqs).txt

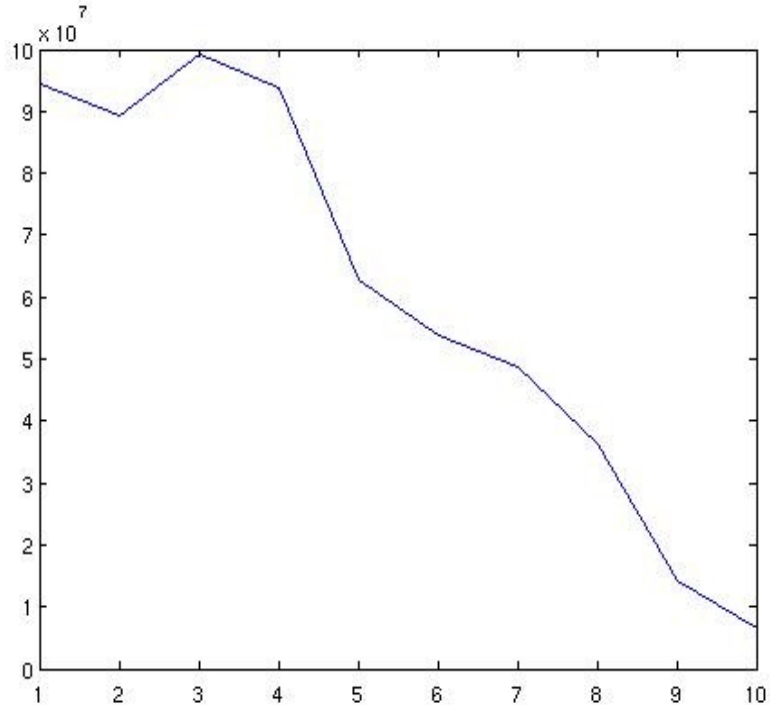
Η εκπαίδευση πραγματοποιήθηκε για διάρκεια 10 εποχών, πλέγμα 9x9 και HMM μίας κατάστασης. Ο Probabilistic Distance Matrix (PDM) μετά το πέρας της εκπαίδευσης απεικονίζεται στην εικόνα 6.1 ενώ τα μέτρα οργάνωσης στην 6.2:



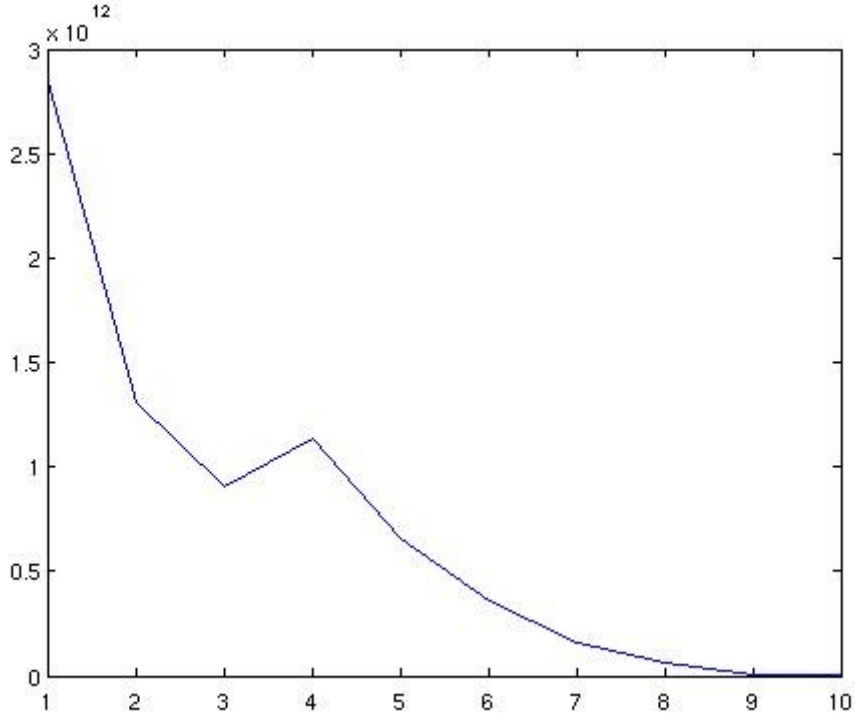
Εικόνα 6.1. PDM για το πρώτο τεχνητό σύνολο δεδομένων. Πάνω αριστερά εδράζεται η πρώτη συστάδα, η δεύτερη άνω δεξιά και η τρίτη στο κάτω κομμάτι του χάρτη. Οι συστάδες έχουν σαφή διαχωρισμό, χωρίς να παρατηρείται κόμβος που να περιέχει στοιχεία από δύο διαφορετικές.



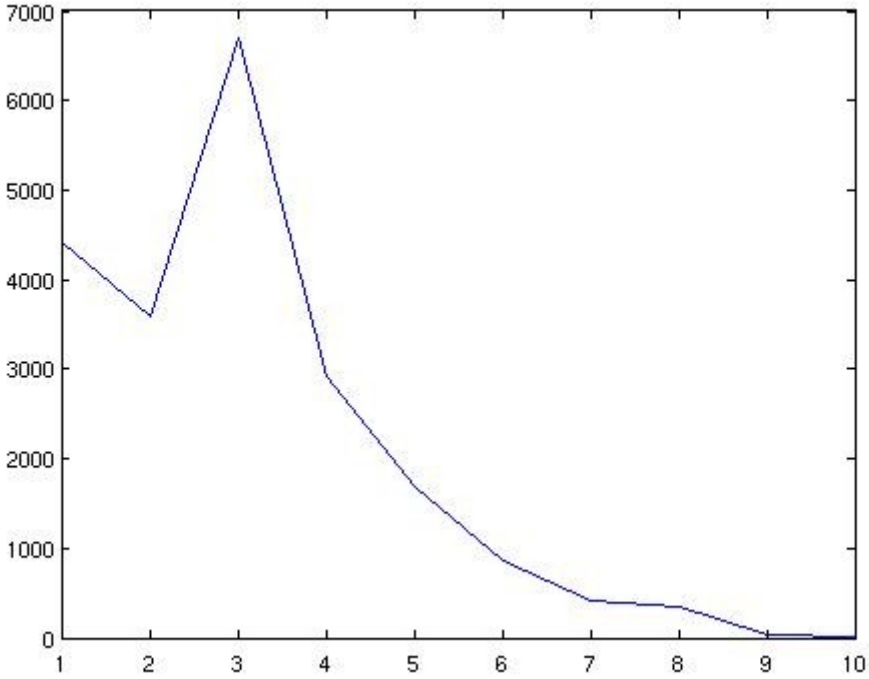
(α) C-measure



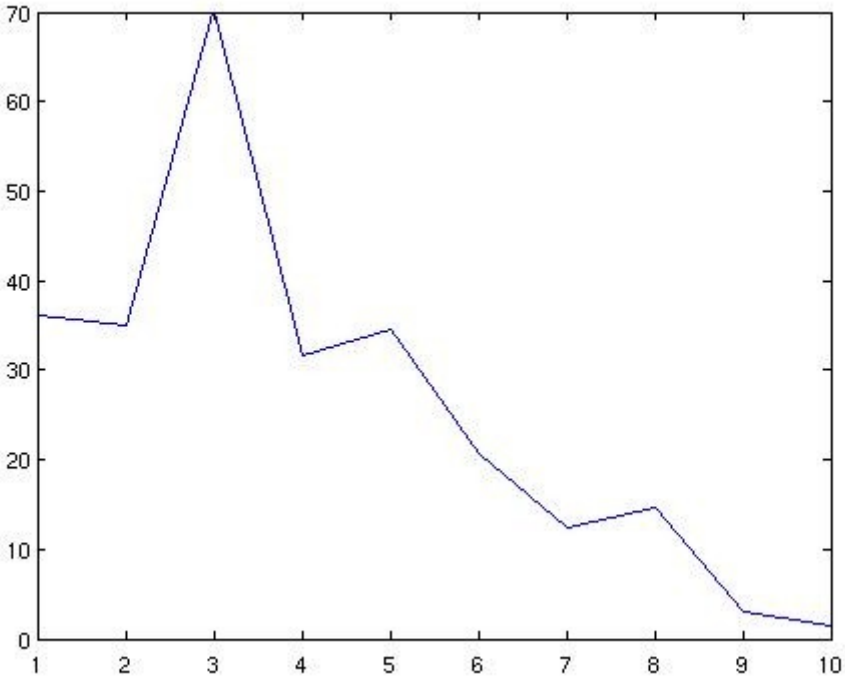
(β) Calinski-Harabasz



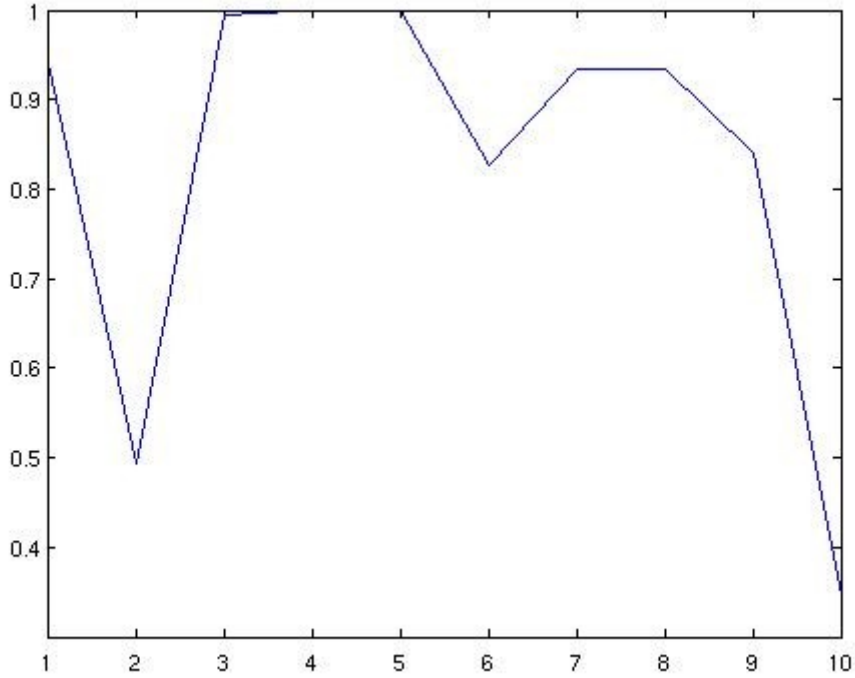
(γ) Demartines-Herault



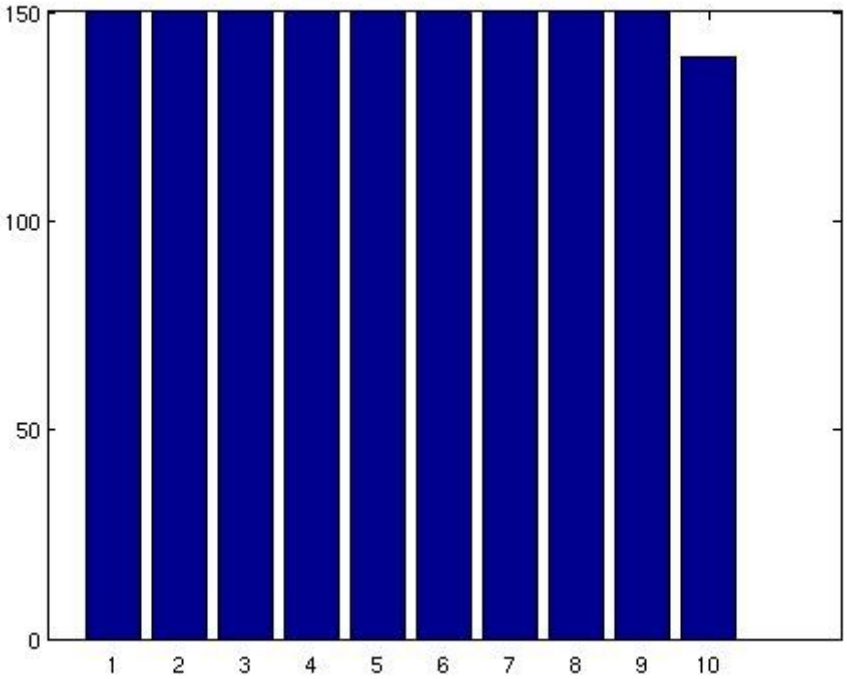
(δ)Distortion measure



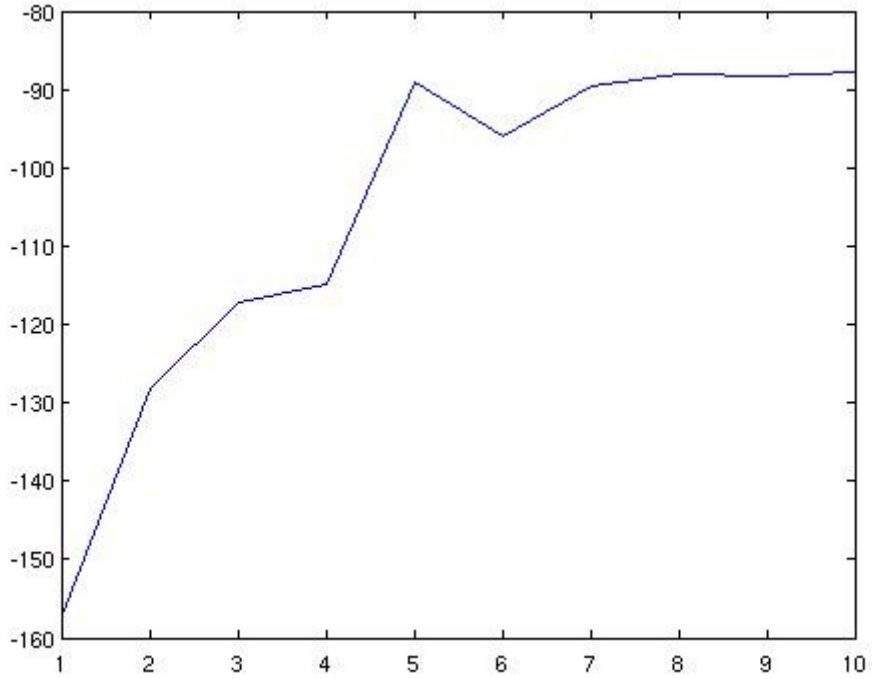
(ε) Quantization error



(ζ) topographic error



(η) ακολουθίες που άλλαξαν κόμβο

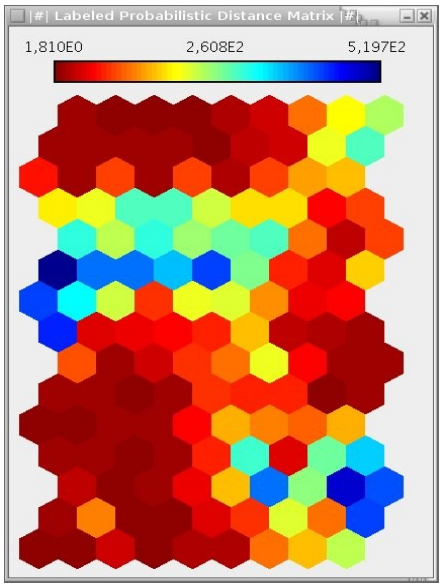


(θ)Εντροπία (σε λογαριθμική κλίμακα)

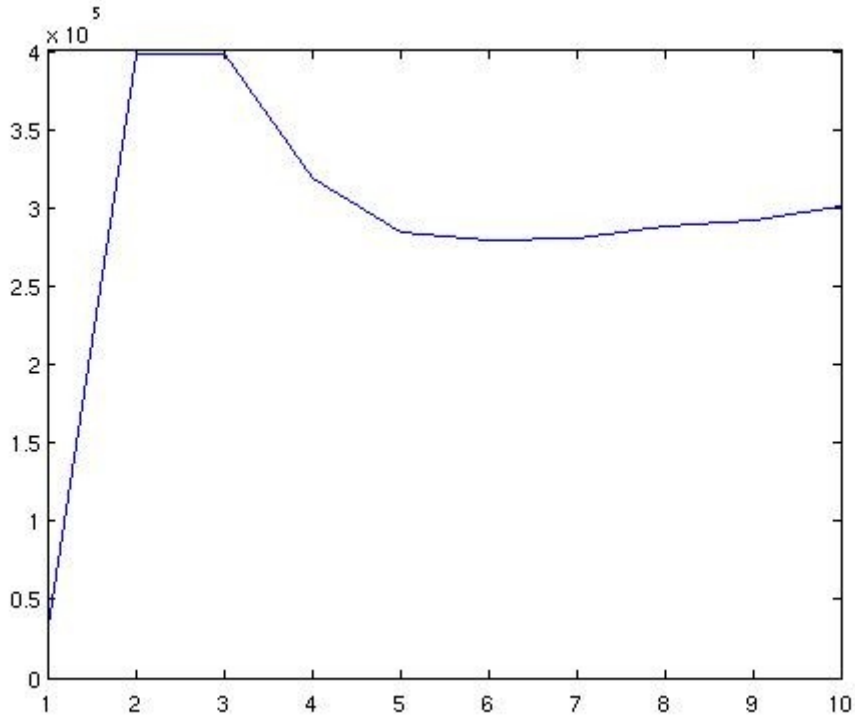
Εικόνα 6.2 Μέτρα οργάνωσης για το πρώτο τεχνητό σύνολο δεδομένων

SecondArtificial(seqs).txt

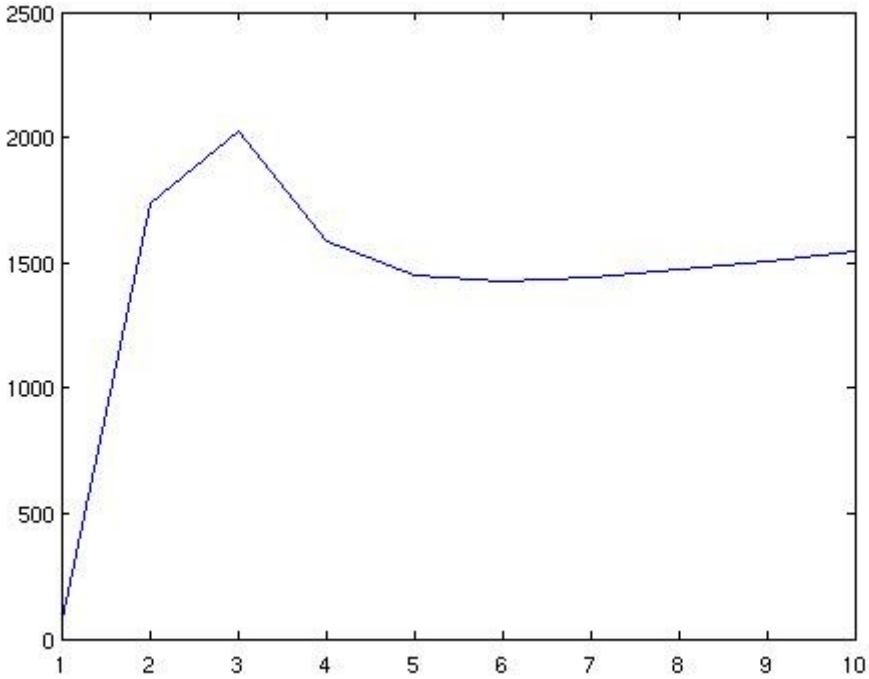
Η εκπαίδευση πραγματοποιήθηκε για διάρκεια 10 εποχών, πλέγμα 5x8 και HMM δέκα καταστάσεων. Ο Probabilistic Distance Matrix (PDM) μετά το πέρας της εκπαίδευσης απεικονίζεται στην εικόνα 6.3 ενώ τα μέτρα οργάνωσης στην 6.4:



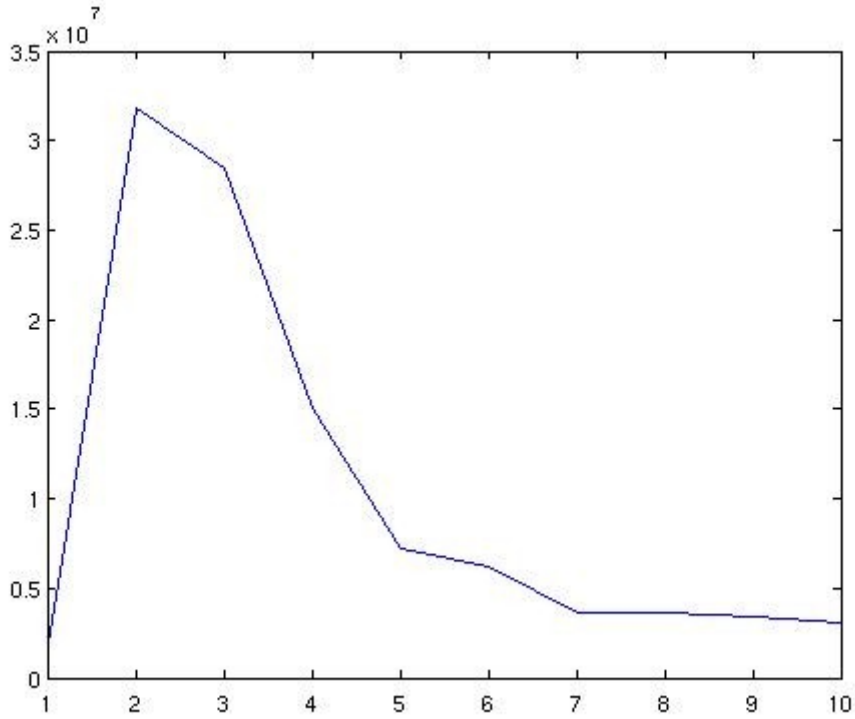
Εικόνα 6.3. PDM για το δεύτερο τεχνητό σύνολο δεδομένων.



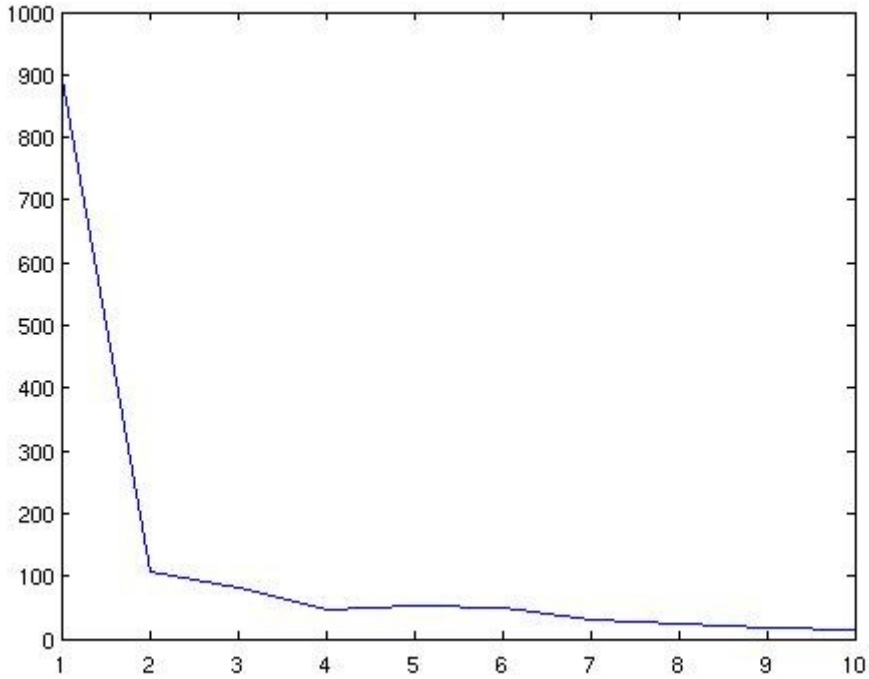
(α) C-measure



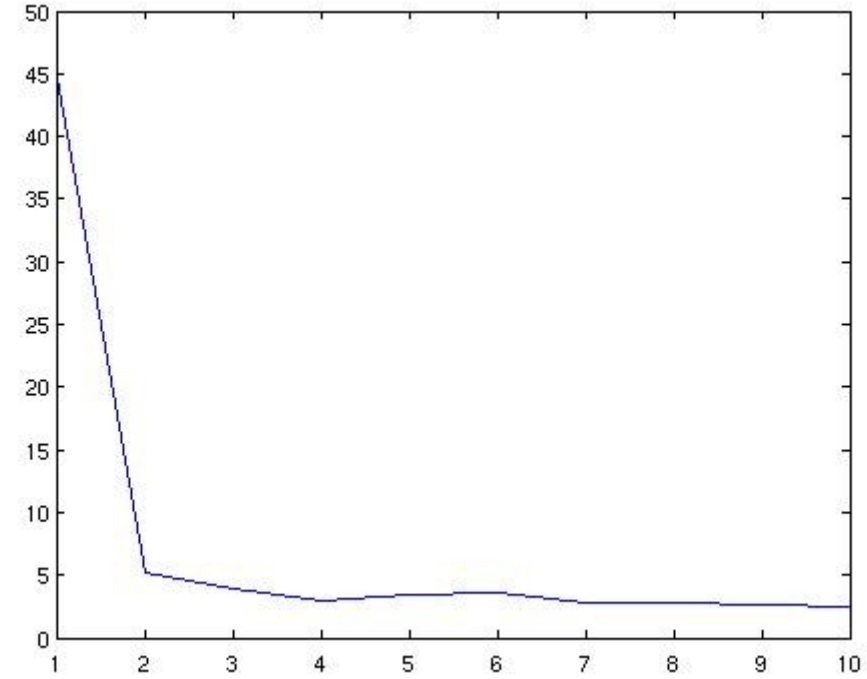
(β) Calinski-Harabasz



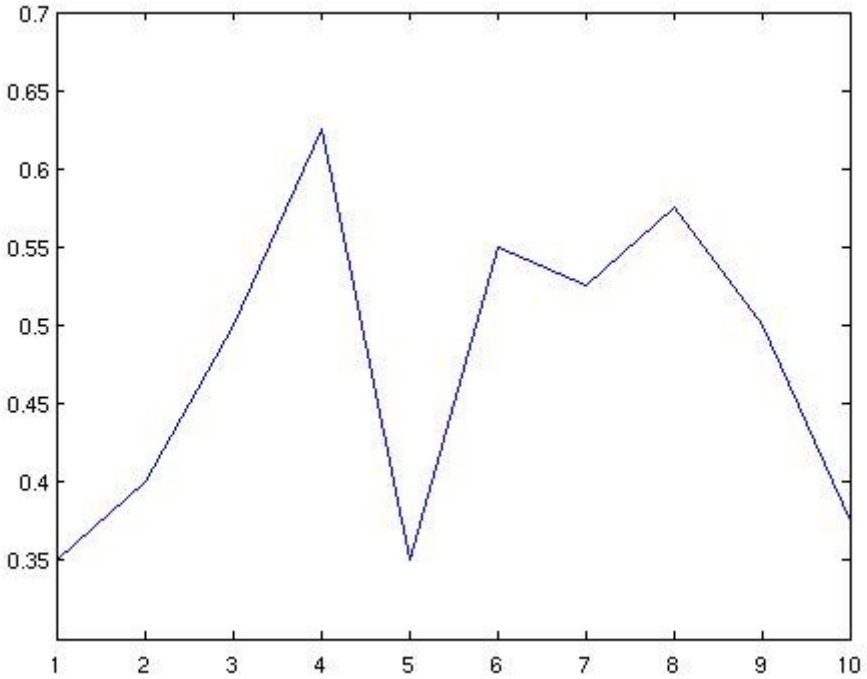
(γ) Demartines-Herault



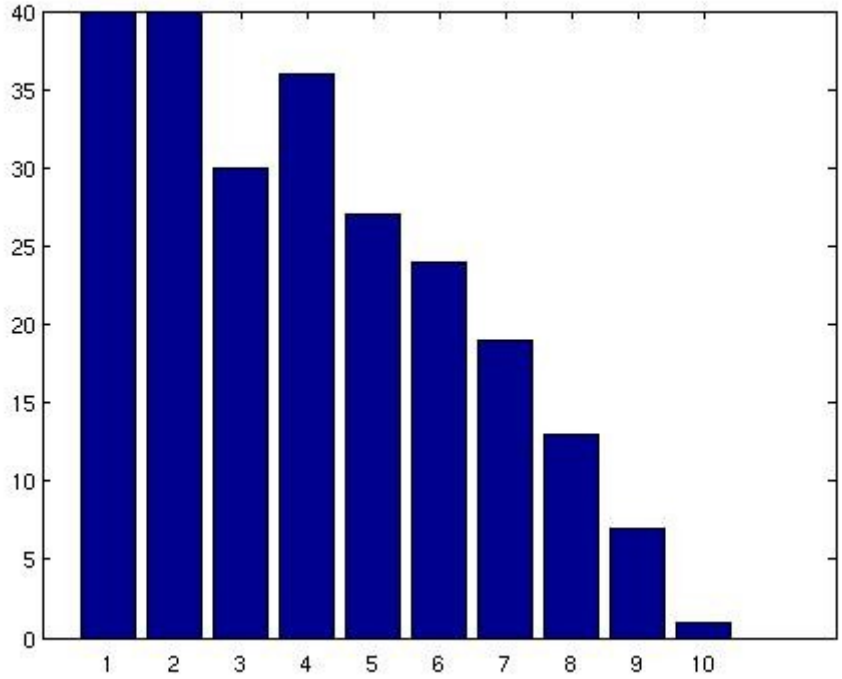
(δ)Distortion measure



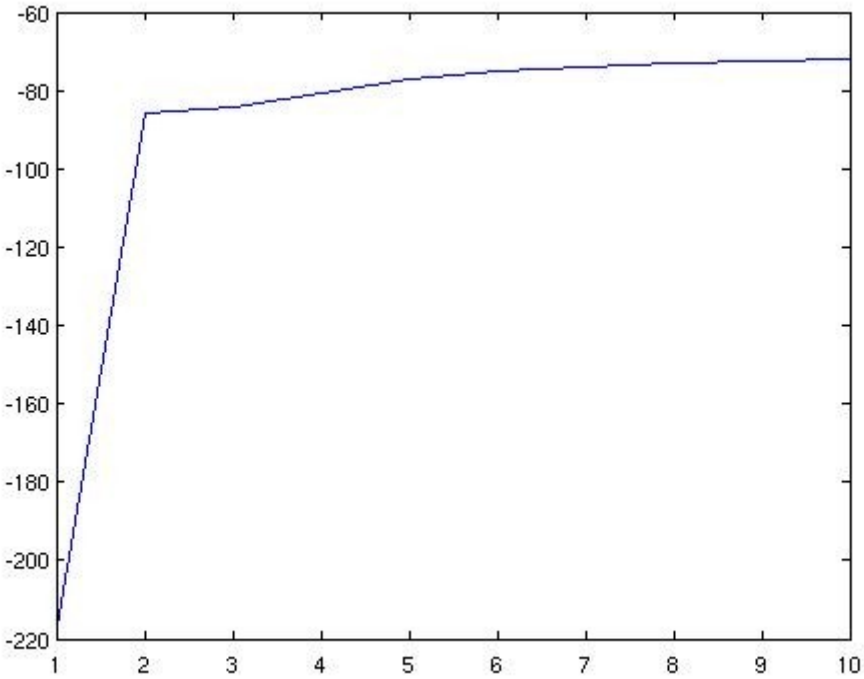
(ε) Quantization error



(ζ) topographic error



(η) ακολουθίες που άλλαξαν κόμβο

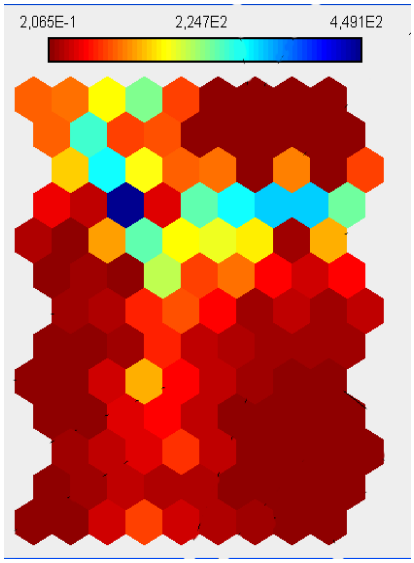


(θ)Εντροπία (σε λογαριθμική κλίμακα)

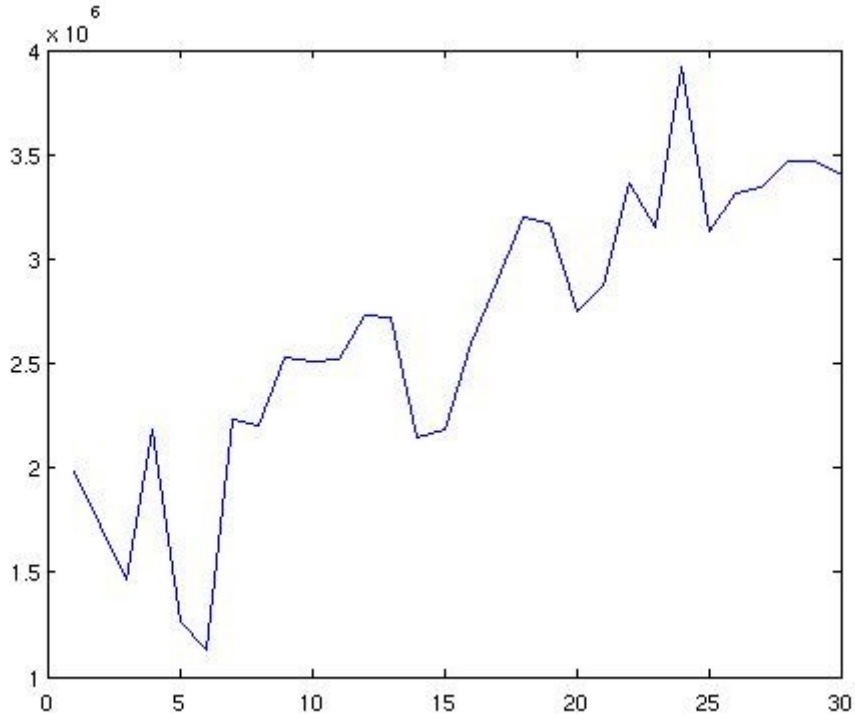
Εικόνα 6.4 Μέτρα οργάνωσης για το δεύτερο τεχνητό σύνολο δεδομένων

ThirdArtificial(seqs).txt

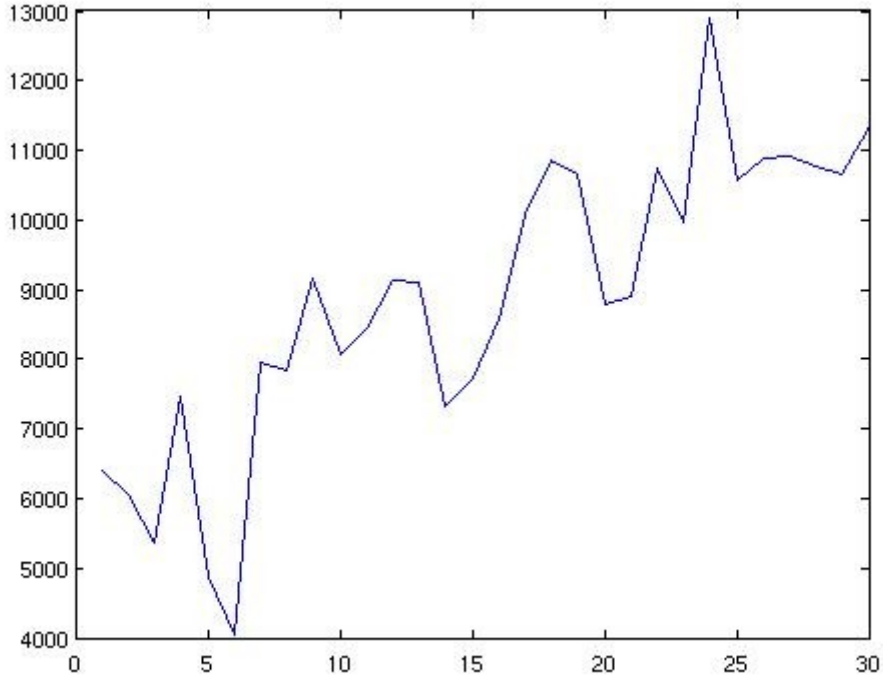
Η εκπαίδευση πραγματοποιήθηκε για διάρκεια 30 εποχών, πλέγμα 5x7 και HMM πέντε καταστάσεων. Ο Probabilistic Distance Matrix (PDM) μετά το πέρας της εκπαίδευσης απεικονίζεται στην εικόνα 6.5 ενώ τα μέτρα οργάνωσης στην 6.6:



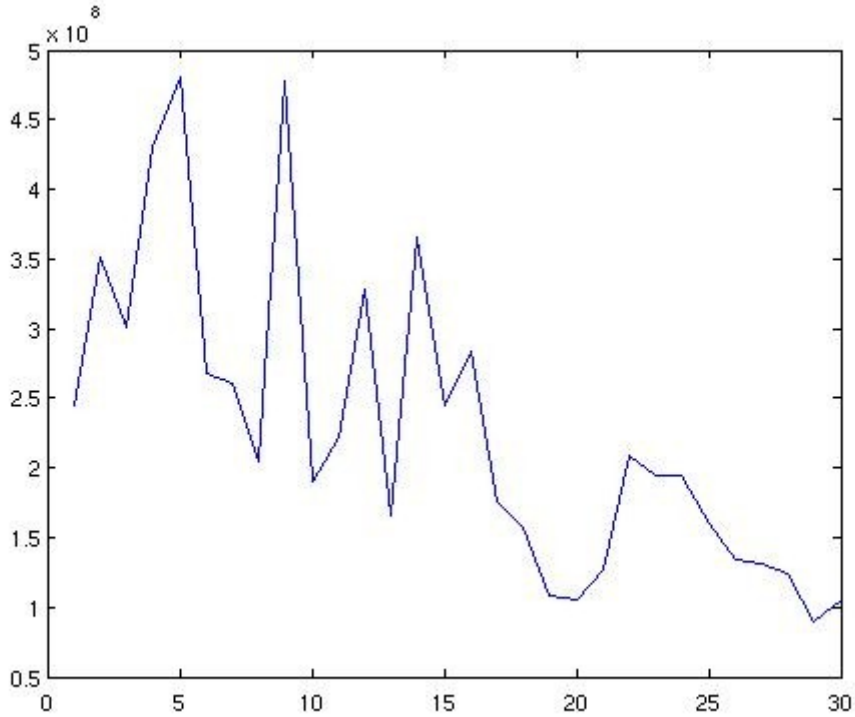
Εικόνα 6.5. PDM για το τρίτο τεχνητό σύνολο δεδομένων.



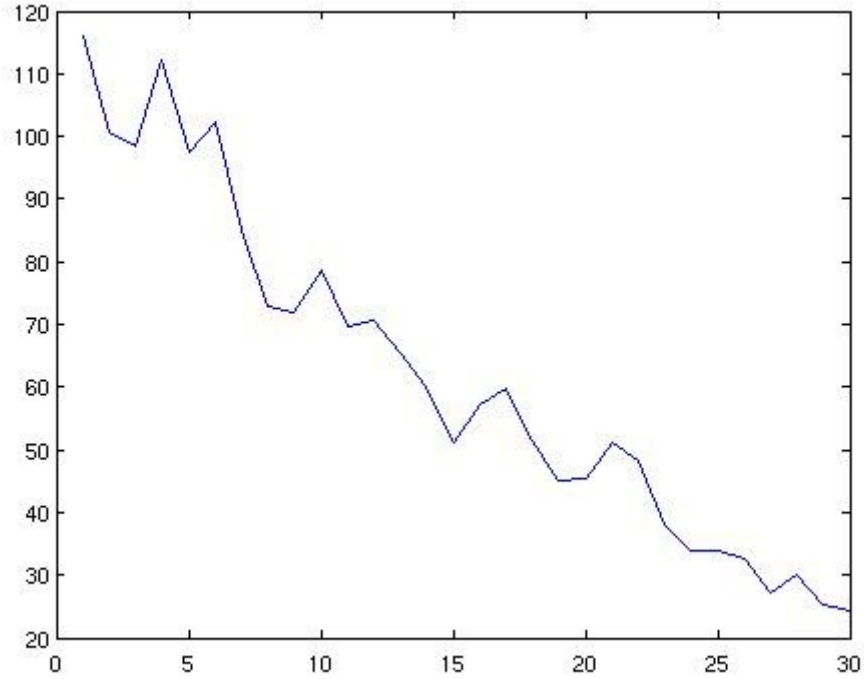
(α) C-measure



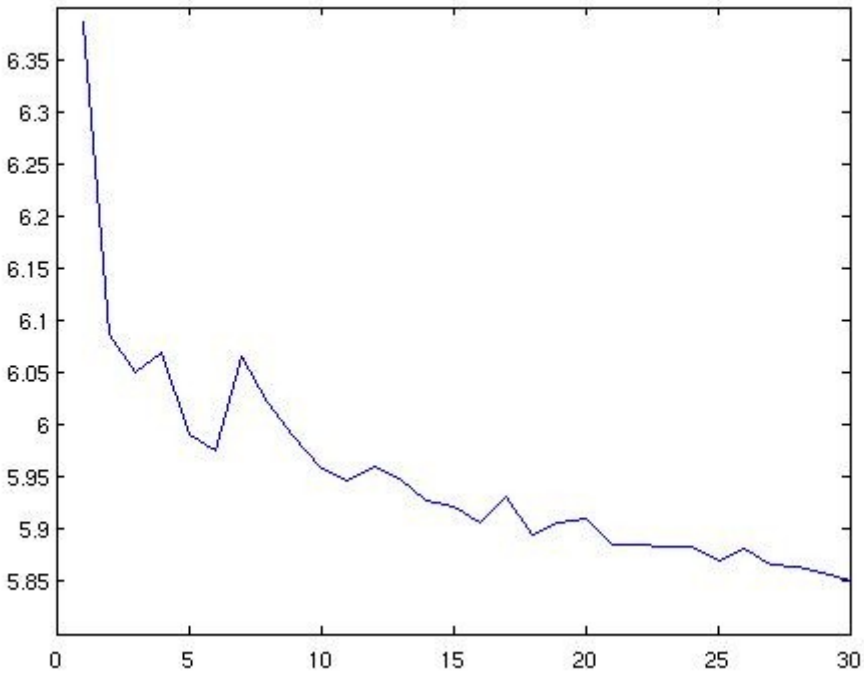
(β) Calinski-Harabasz



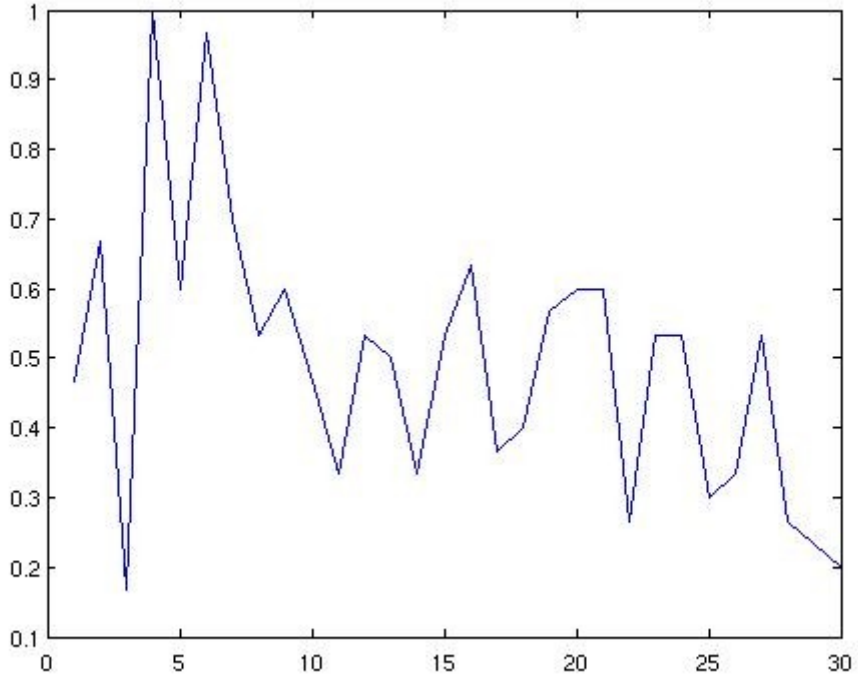
(γ) Demartines-Herault



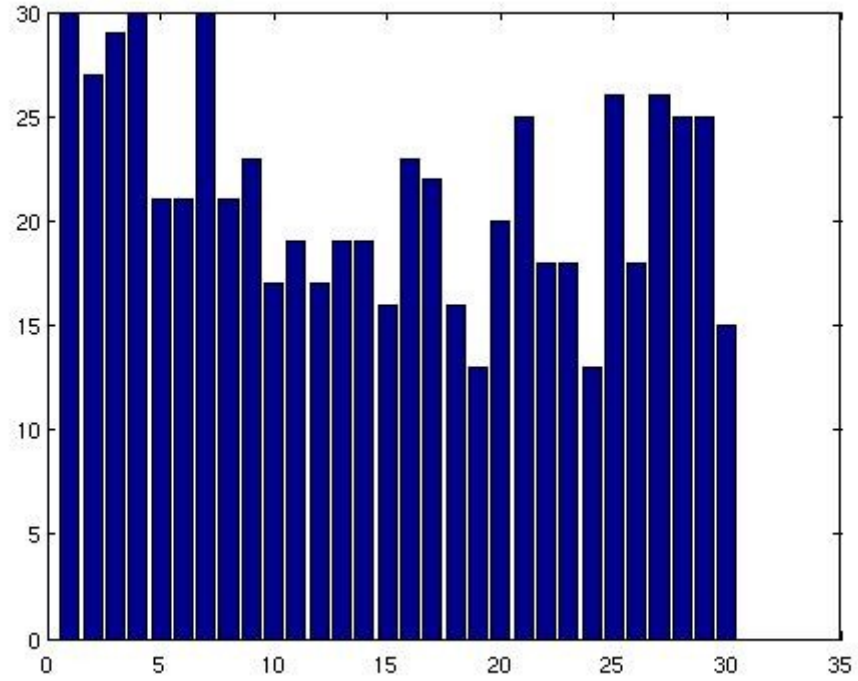
(δ)Distortion measure



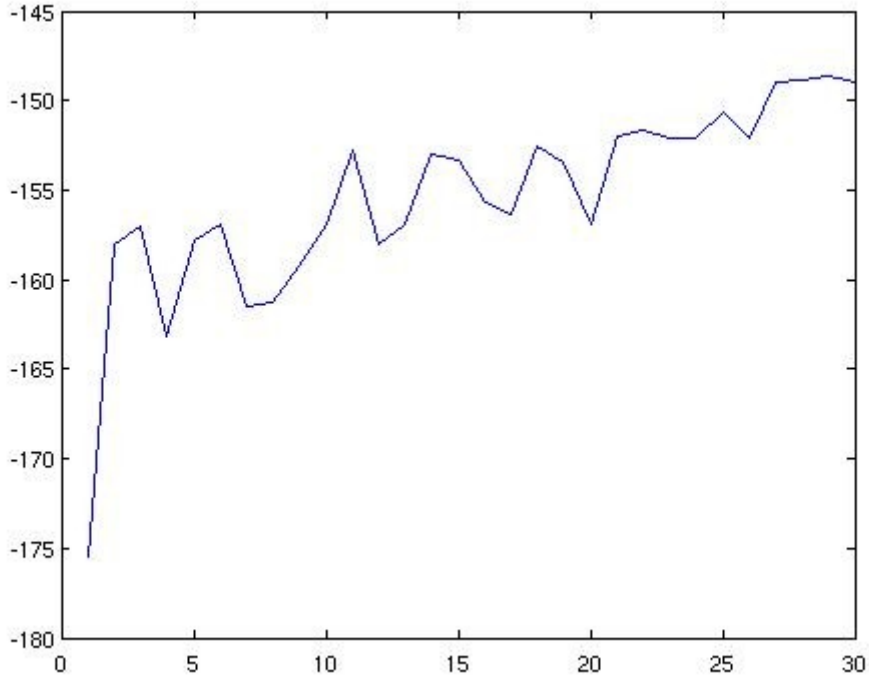
(ε) Quantization error



(ζ) topographic error



(η) ακολουθίες που άλλαξαν κόμβο



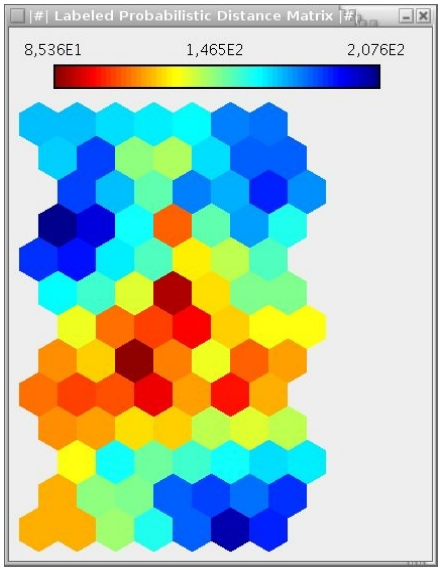
(θ)Εντροπία (σε λογαριθμική κλίμακα)

Εικόνα 6.4 Μέτρα οργάνωσης για το δεύτερο τεχνητό σύνολο δεδομένων

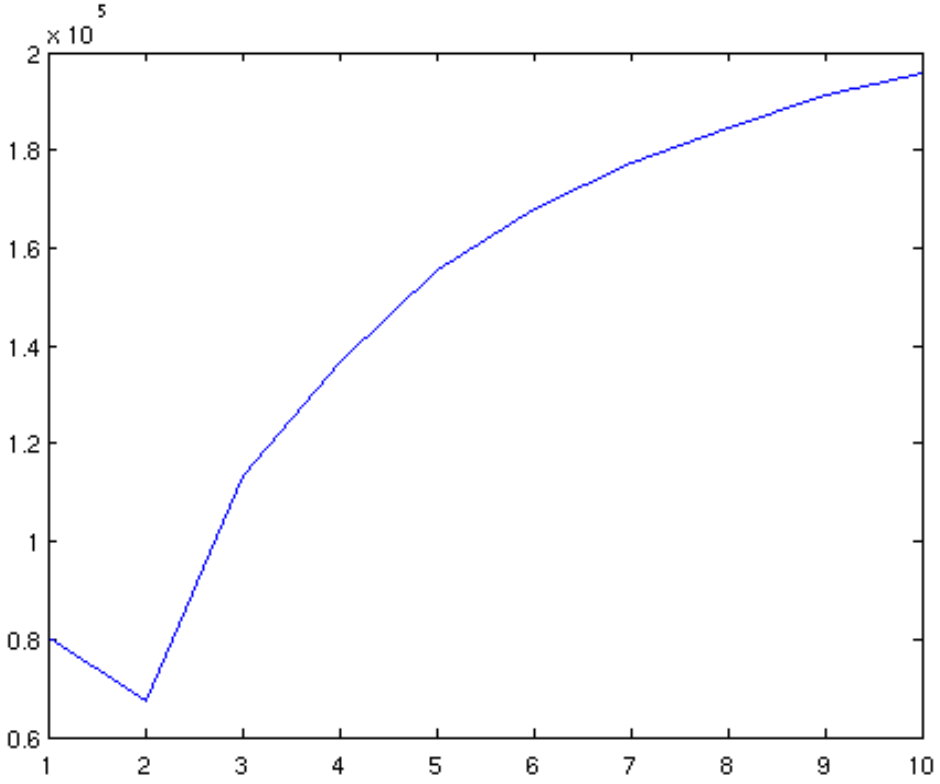
6.3.1.2 Πραγματικά σύνολα δεδομένων

Globins 1

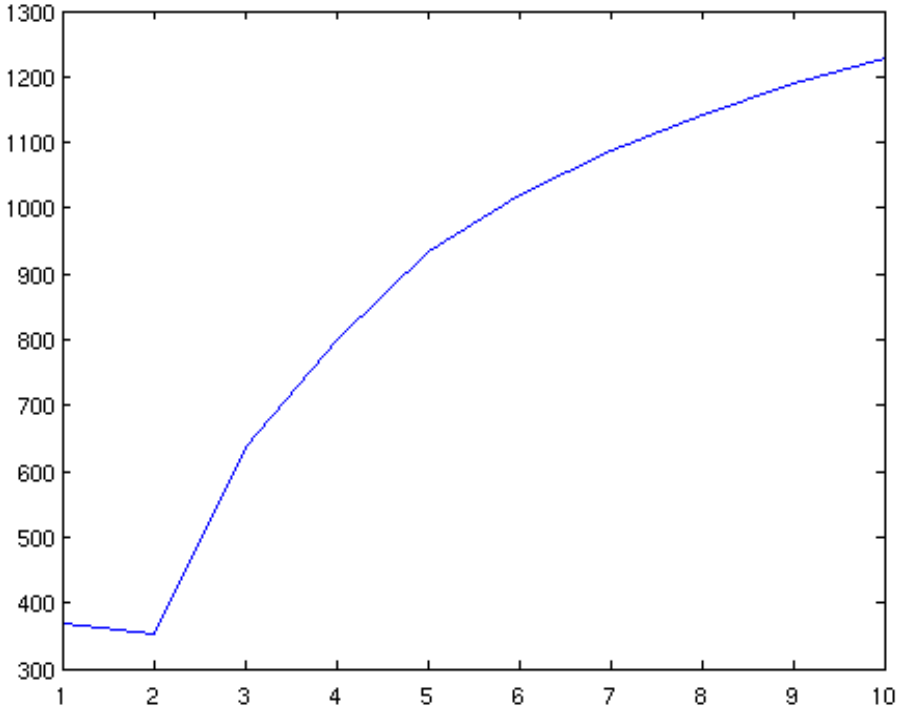
Η εκπαίδευση πραγματοποιήθηκε για διάρκεια 10 εποχών, πλέγμα 7x4 και HMM έντεκα καταστάσεων. Ο Probabilistic Distance Matrix (PDM) μετά το πέρας της εκπαίδευσης απεικονίζεται στην εικόνα 6.7 ενώ τα μέτρα οργάνωσης στην 6.8:



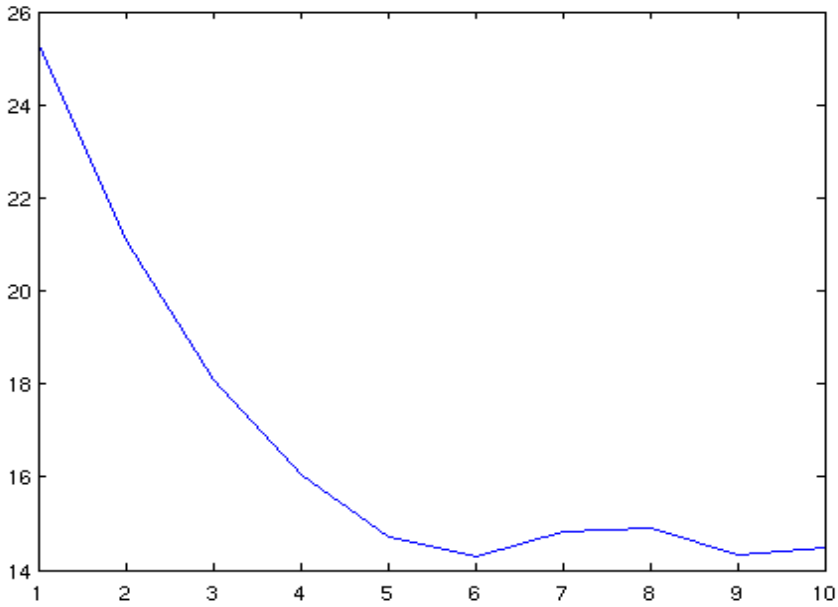
Εικόνα 6.7. PDM για το πρώτο πραγματικό σύνολο δεδομένων. Πάνω αριστερά εντοπίζεται η συστάδα με τα A-globins, πάνω δεξιά αυτή με τα B-globins και εμφανώς διακριτή κάτω δεξιά βρίσκεται η συστάδα των myoglobins.



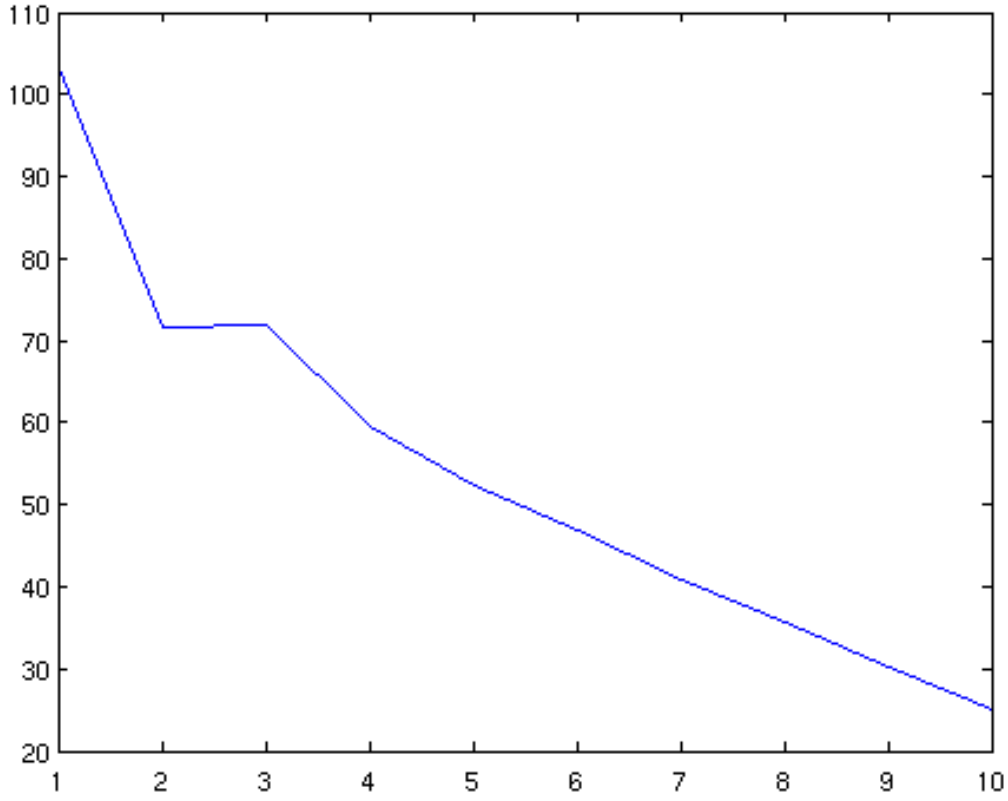
(α) C-measure



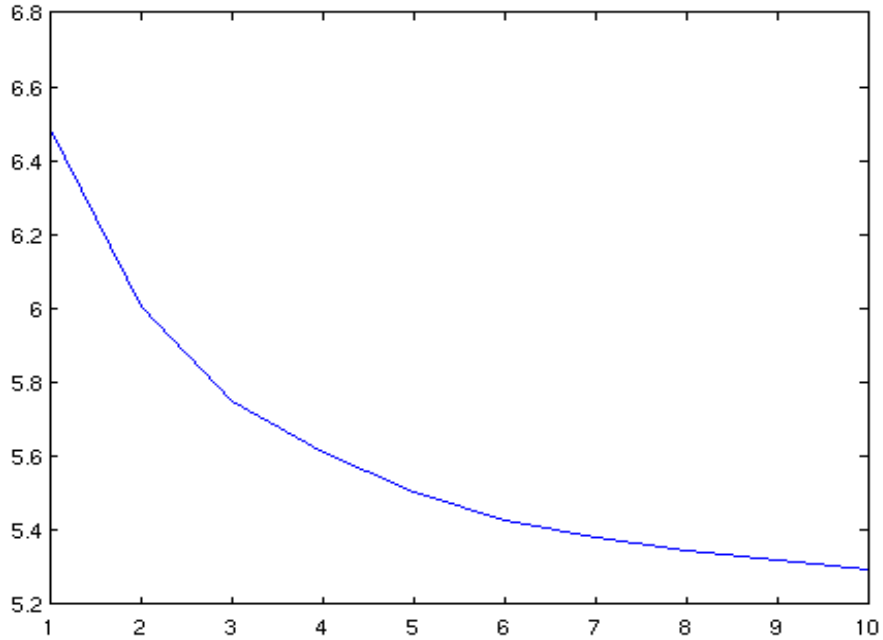
(β) Calinski-Harabasz



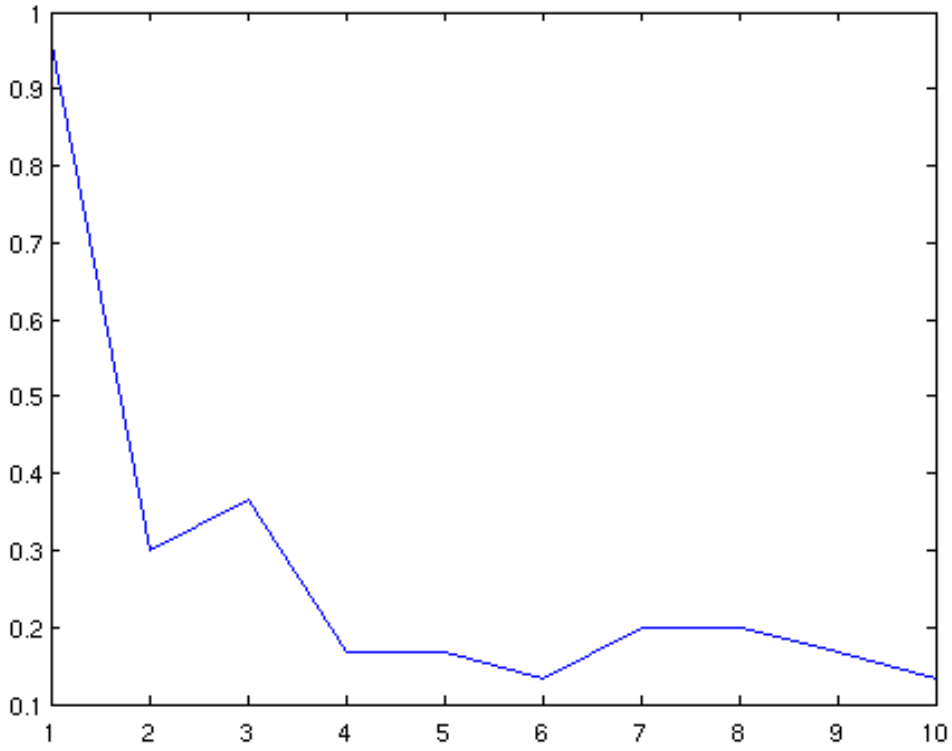
(γ) Demartines-Herault



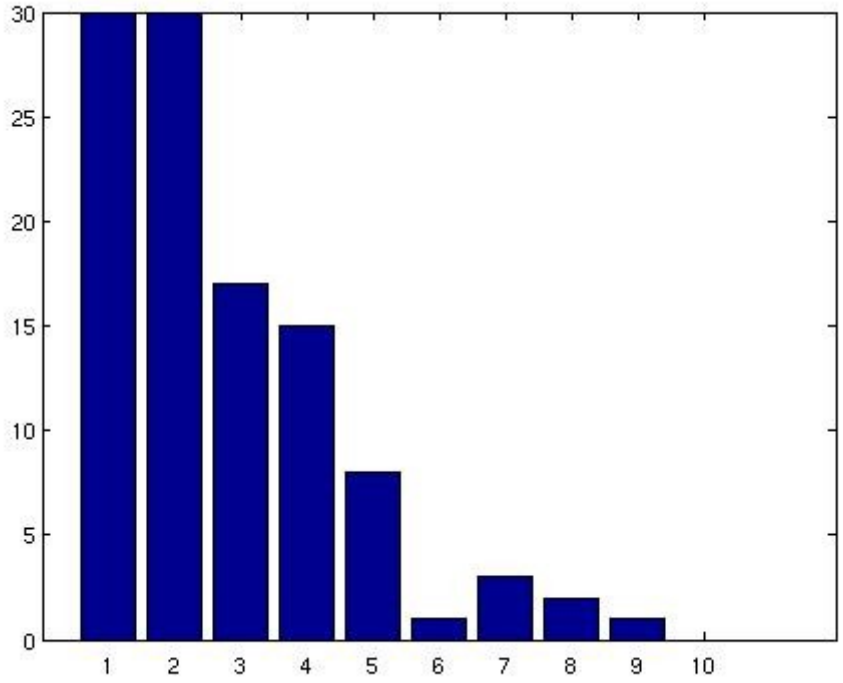
(δ)Distortion measure



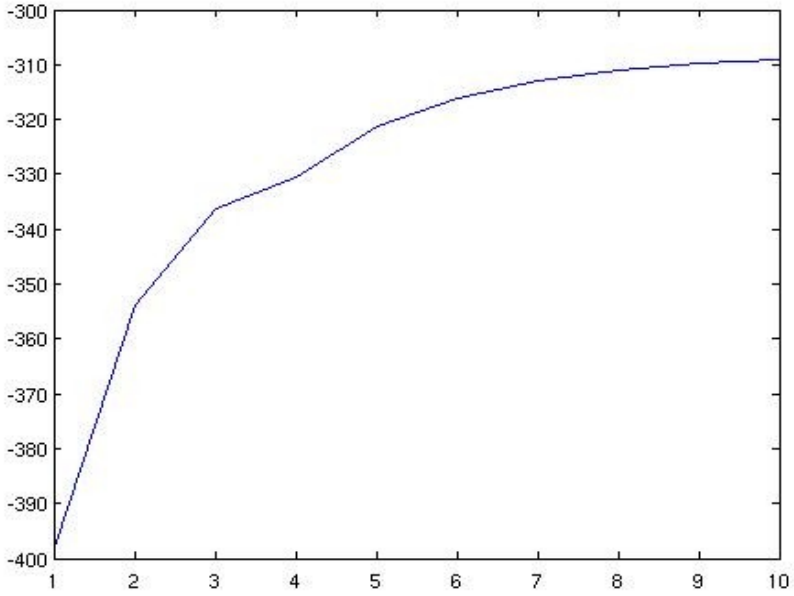
(ε) Quantization error



(ζ) topographic error



(η) ακολουθίες που άλλαξαν κόμβο

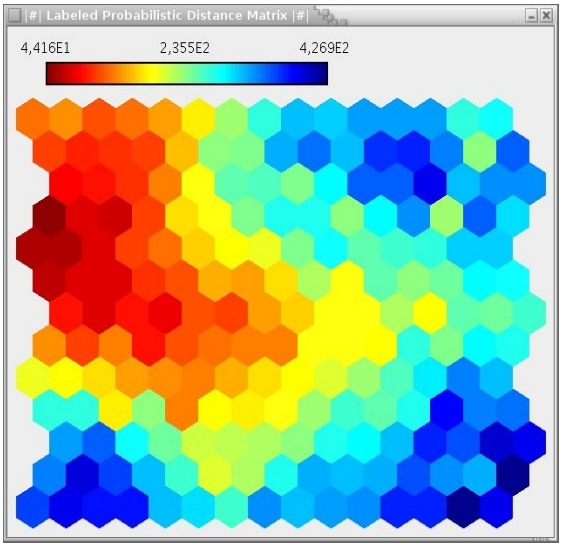


(θ)Εντροπία (σε λογαριθμική κλίμακα)

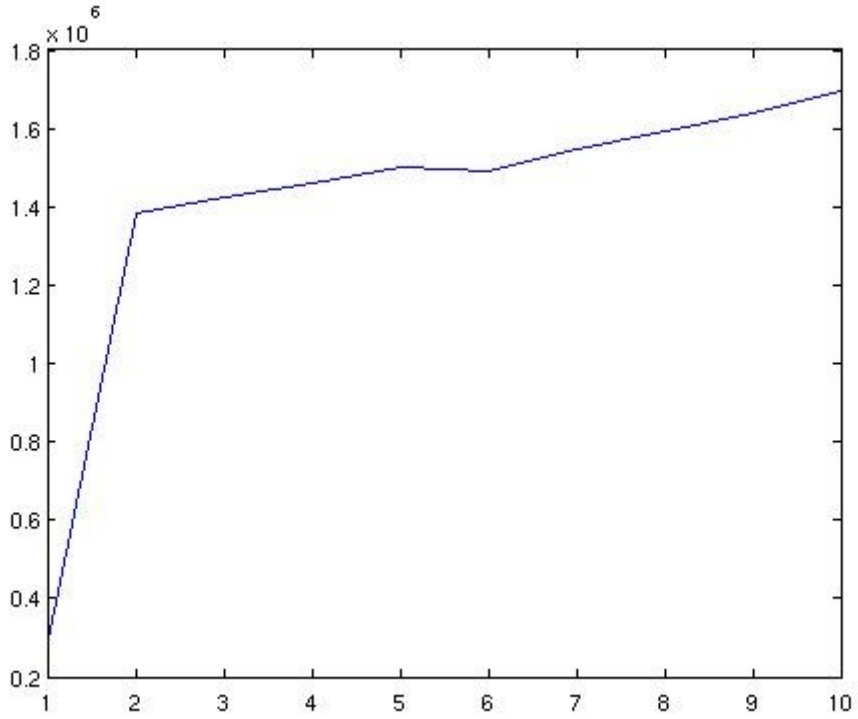
Εικόνα 6.8 Μέτρα οργάνωσης για το πρώτο πραγματικό σύνολο δεδομένων

Globins 2

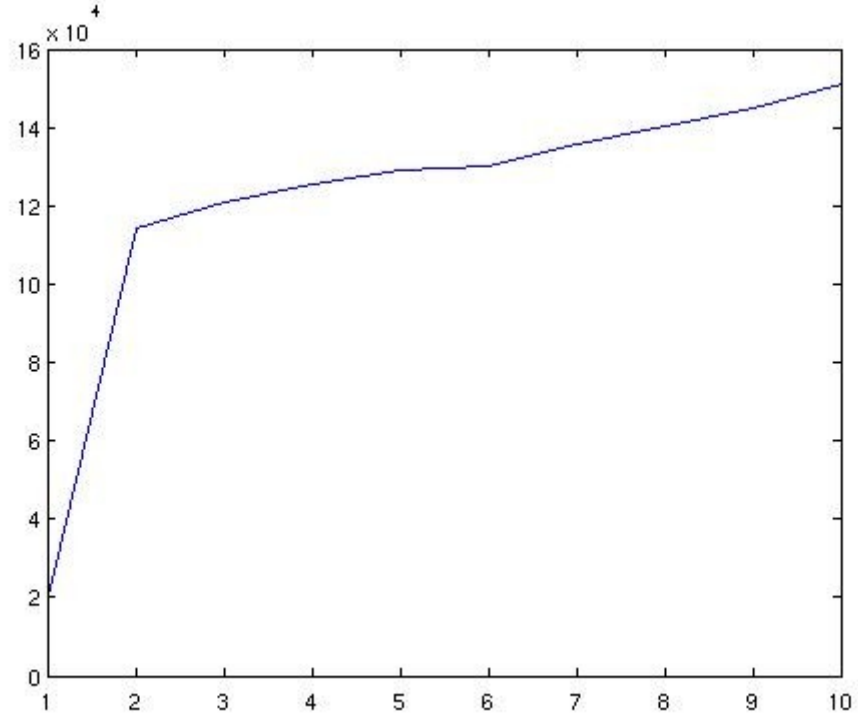
Η εκπαίδευση πραγματοποιήθηκε για διάρκεια 10 εποχών, πλέγμα 7x8 και HMM έντεκα καταστάσεων. Ο Probabilistic Distance Matrix (PDM) μετά το πέρας της εκπαίδευσης απεικονίζεται στην εικόνα 6.9 ενώ τα μέτρα οργάνωσης στην 6.10:



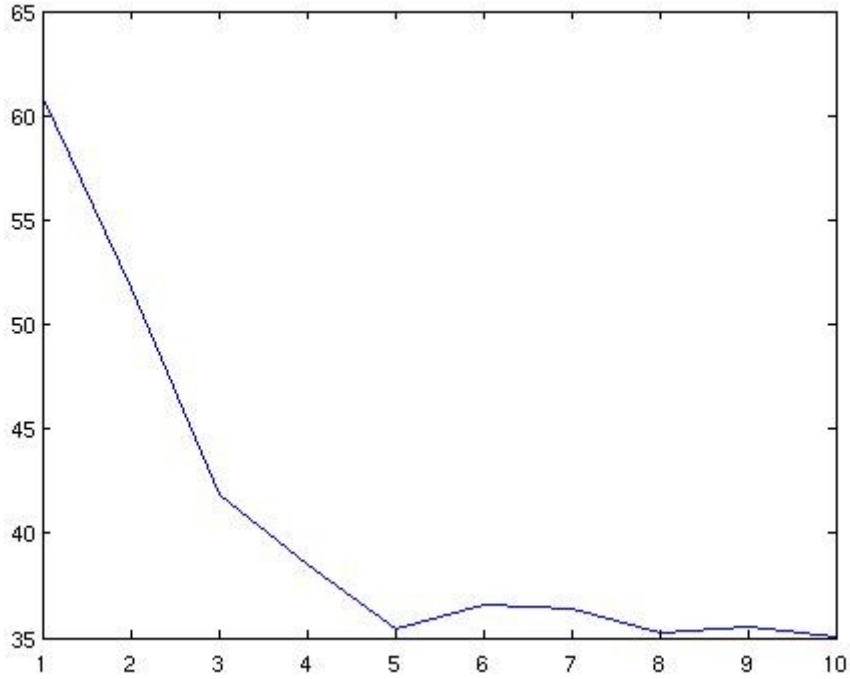
Εικόνα 6.9. PDM για το δεύτερο πραγματικό σύνολο δεδομένων. A-globins και B-globins εικονίζονται αντίστοιχα στο άνω δεξιά και κάτω δεξιά μέρος του χάρτη σε διακριτές συστάδες, ενώ στην κάτω αριστερή πλευρά εντοπίζεται η συστάδα των myoglobins.



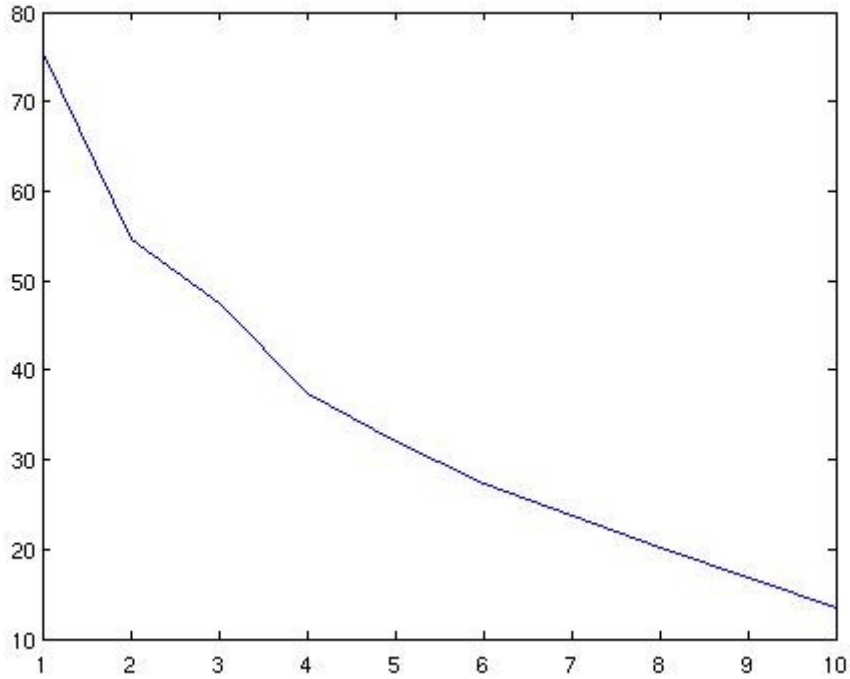
(α) C-measure



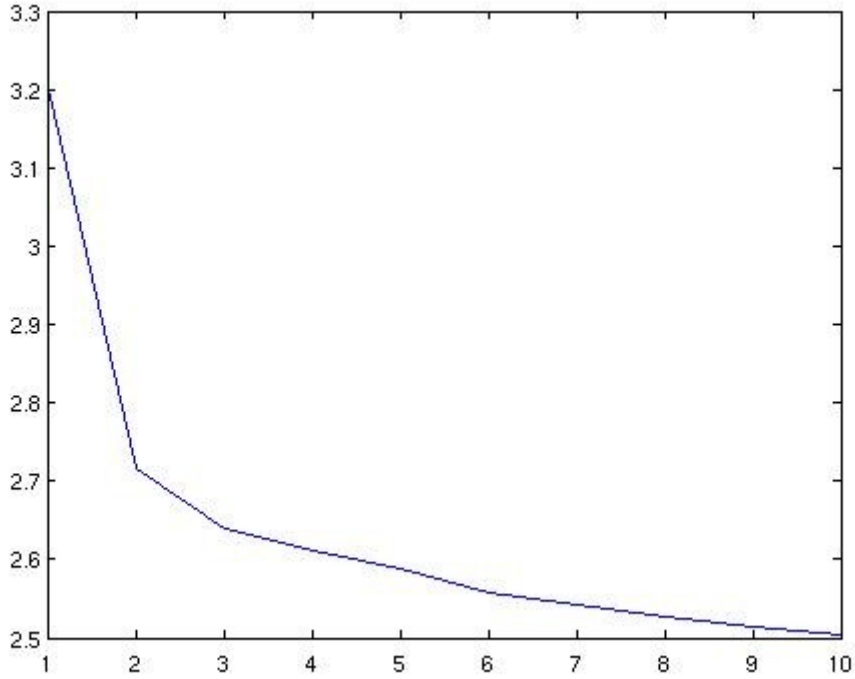
(β) Calinski-Harabasz



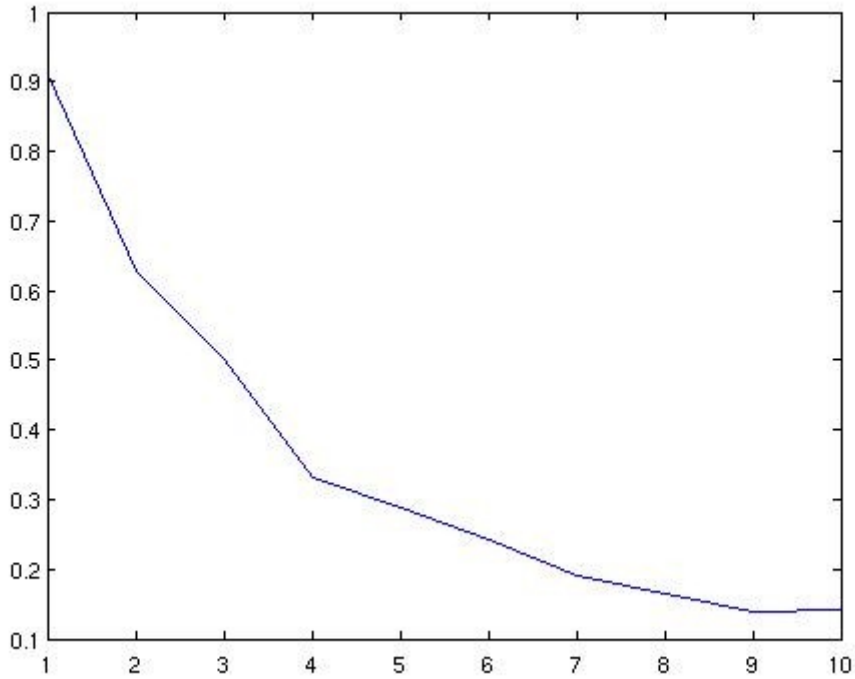
(γ) Demartines-Herault



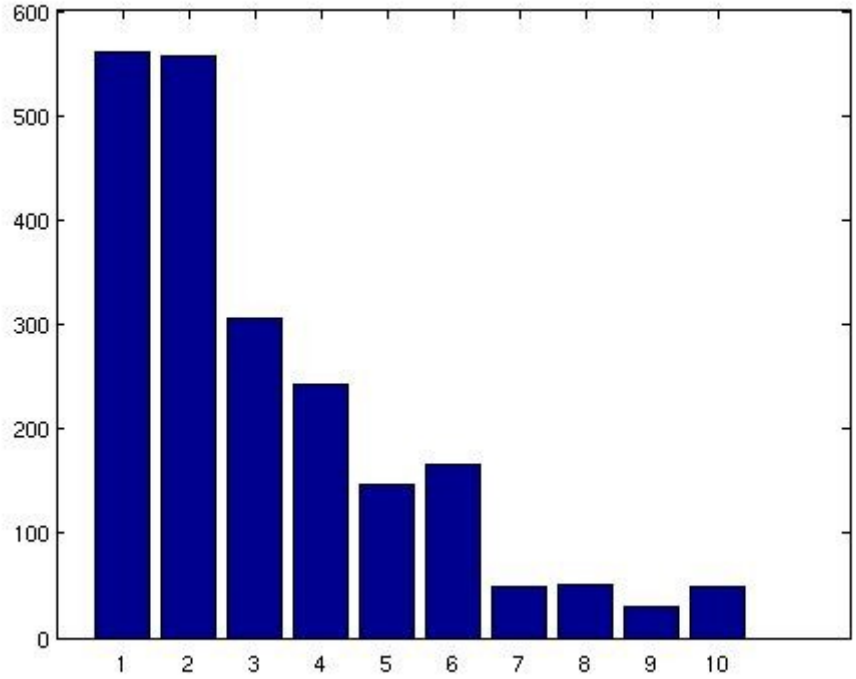
(δ)Distortion measure



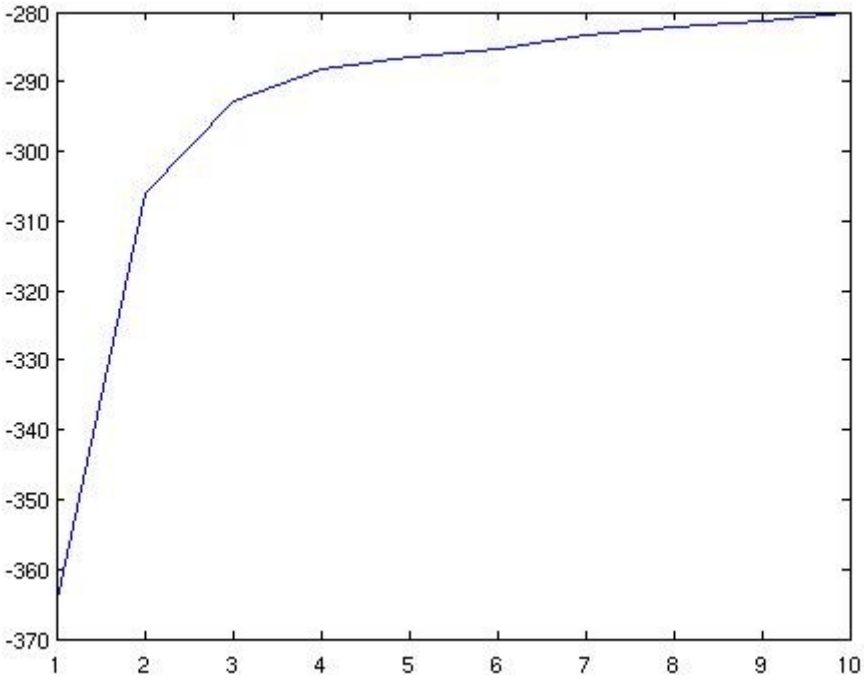
(ε) Quantization error



(ζ) topographic error



(η) ακολουθίες που άλλαξαν κόμβο



(θ)Εντροπία (σε λογαριθμική κλίμακα)

Εικόνα 6.10 Μέτρα οργάνωσης για το δεύτερο πραγματικό σύνολο δεδομένων

6.3.1.3 Συμπεράσματα

Μελετώντας τα παραπάνω γραφήματα για τα διάφορα μέτρα οργάνωσης και για διαφορετικά σύνολα δεδομένων μπορούμε εν γένει να καταλήξουμε σε κάποια συμπεράσματα ως προς την απόδοσή τους.

Αρχικά, σε ότι αφορά το μέτρο *C* για το οποίο αναμένουμε να αυξάνεται κατά τη διάρκεια της εκπαίδευσης (βλ. Κεφάλαιο 5), παρατηρούμε ότι συμπεριφέρεται ικανοποιητικά καλά στην περίπτωση των πραγματικών συνόλων δεδομένων, καθώς και για την περίπτωση του δεύτερου και τρίτου τεχνητού συνόλου δεδομένων. Ειδικά, για το δεύτερο σύνολο δεδομένων εφιστούμε την προσοχή στην απότομη άνοδο για τις πρώτες δύο εποχές, που συνοδεύεται από μια μείωση του ρυθμού ανόδου. Η εξέλιξη αυτή απεικονίζει προφανώς την γρήγορη, αρχικά πιο χονδρική (rough) διάταξη των δεδομένων εισόδου στον χάρτη, και την μετέπειτα ομαλότερη τακτοποίησή τους (tuning). Βλέπουμε συνεπώς, πώς το μέτρο C είναι σε θέση να ποσοτικοποιήσει με ικανοποιητικό τρόπο την εξέλιξη της ποιότητας της διάταξης στο χρόνο. Επίσης, είναι ιδιαίτερα ενθαρρυντική η παρατηρούμενη συμπεριφορά σε ότι αφορά τα πραγματικά σύνολα δεδομένων όπου η μονοτονία του μέτρου είναι σχεδόν τέλεια. Η παρατήρηση αυτή, για τα πραγματικά datasets ισχύει για όλα τα μέτρα οργάνωσης, σε μικρότερο ή μεγαλύτερο βαθμό. Στην περίπτωση του πρώτου συνόλου δεδομένων, το μέτρο C έχει συμπεριφορά αντίθετη από την αναμενόμενη, παρουσιάζοντας φθίνουσες τιμές στο χρόνο. Παρατηρώντας ότι όλα τα άλλα κριτήρια, εκτός του Calinski-Harabasz, συμπεριφέρονται επιθυμητά αναζητούμε τους λόγους που προξενούν την ανεπιθύμητη συμπεριφορά των εν λόγω δύο μέτρων οργάνωσης. Παρατηρώντας το κριτήριο Calinski-Harabasz, βλέπουμε ότι η τιμή του δίνεται από τον λόγο *BGSS/WGSS* (βλέπε Κεφάλαιο 5) όπου *BGSS* είναι το άθροισμα των τετραγώνων των αποστάσεων μεταξύ των νευρώνων και *WGSS* το άθροισμα των τετραγώνων των αποστάσεων μεταξύ κάθε νευρώνα και των προτύπων που αυτός περιγράφει. Το *WGSS* αποτελεί το σφάλμα κβαντισμού, του οποίου η τιμή μειώνεται με τον χρόνο. Αντίθετα, η τιμή του *BGSS* δεν αυξάνεται, όπως αναμένεται, στον χρόνο αλλά μειώνεται. Η παρατήρηση αυτή μας οδηγεί στο να συμπεράνουμε πως στην περίπτωση αυτού του συνόλου δεδομένων δεν λαμβάνουμε επιθυμητή συμπεριφορά όσον αφορά τα μέτρα *C* και Calinski-Harabasz επειδή οι αποστάσεις μεταξύ των κόμβων στο σύνολό τους μειώνονται (αντί να αυξάνονται όπως στις άλλες περιπτώσεις). Πιθανοί λόγοι που συμβαίνει αυτό μπορεί να είναι το γεγονός ότι τα δεδομένα εισόδου έχουν μικρή διασπορά (μέσα σε κάθε μια από τις τρεις συστάδες) και συνεπώς δεν χρειάζονται τόσο μεγάλο πλέγμα προκειμένου να περιγραφεί η δομή τους. Με άλλα λόγια, ένας μεγάλος αριθμός κόμβων καλείται να περιγράψει δεδομένα πολύ όμοια μεταξύ τους, οπότε και το άθροισμα των αποστάσεων μεταξύ των κόμβων αναμένεται να μειωθεί στη διάρκεια της εκπαίδευσης (αφού αποκτούν τιμές πολύ όμοιες).

Στην περίπτωση του κριτηρίου Calinski-Harabasz, παρατηρείται μια παρόμοια συμπεριφορά με αυτή του μέτρου *C*, συνεπώς τα συμπεράσματα θα μπορούσαμε να πούμε πως είναι τα ίδια. Συχνά, οι καμπύλες για το ίδιο σύνολο δεδομένων έχουν ίδια μορφή με αυτή του *C*, παρουσιάζοντας κοινά τοπικά ακρότατα. Το συμπέρασμα αυτό είναι αρκετά ενδιαφέρον, ιδίως αν λάβουμε υπόψιν ότι ο τρόπος υπολογισμού των δύο μέτρων είναι εντελώς διαφορετικός. Σημειώνεται πάντως, ότι το κριτήριο Calinski Harabasz απαιτεί σαφώς περισσότερο χρόνο για τον υπολογισμό του από ότι το μέτρο C.

Το μέτρο των Demartines και Herault αναμένεται να φθίνει κατά τη διάρκεια της εκπαίδευσης. Αν και αρκετά ευαίσθητο, εφόσον επιχειρείται μια επακριβής αντιστοίχηση των αποστάσεων στους χώρους εισόδου και εξόδου (βλ. Κεφάλαιο 5), εντούτοις επιτυγχάνει να περιγράψει την διαδικασία της εκπαίδευσης σε όλες τις περιπτώσεις συνόλων δεδομένων που εξετάσθηκαν. Ενδιαφέρον παρουσιάζει η περίπτωση του δευτέρου τεχνητού συνόλου

δεδομένων. Η τιμή του μέτρου αυξάνεται στις 2 πρώτες εποχές, για να ακολουθήσει στη συνέχεια μονότονα φθίνουσα πορεία. Η αρχική αύξηση αυτή μπορεί να ερμηνευθεί από δύο σκοπιές: σε ότι αφορά τη διαδικασία διάταξης, για τις πρώτες δύο εποχές, όπου ο ρυθμός εκμάθησης έχει μεγάλη τιμή, παρατηρούνται μεγάλες αλλαγές στη διάταξη σε κάθε βήμα κάτι που προκαλεί μια αρχική αστάθεια, ενώ στη συνέχεια, αφού έχει επιτευχθεί η χονδρική διάταξη, η εκπαίδευση πραγματοποιείται ομαλότερα (tuning phase) με χαμηλότερους ρυθμούς μάθησης, κάτι που οδηγεί στην ομαλή, φθίνουσα (όπως αναμένεται) συμπεριφορά του μέτρου Demartines Herault. Από μια δεύτερη σκοπιά, γνωρίζοντας ότι η συνάρτηση αυτή εισήχθη αρχικά από τους παραπάνω ερευνητές ως μια συνάρτηση ενέργειας, μπορούμε να δούμε πως αρχικά παρατηρείται μια αύξηση της ενέργειας (αστάθεια στη διάταξη του χάρτη) και στη συνέχεια μια μείωση, καθώς οδηγούμαστε προς σταθεροποίηση-βελτίωση της διάταξης. Σε συνάρτηση με τα όσα ειπώθηκαν παραπάνω για το μέτρο C και το κριτήριο Calinski Harabasz γίνεται εύκολα αντιληπτό πως τα τρία αυτά μέτρα επιτυγχάνουν να μας δώσουν την ίδια πληροφορία σχετικά με την εξέλιξη της εκπαίδευσης (ordering phase-tuning phase) μέσα από διαφορετικές συμπεριφορές των καμπύλων τους. Στις υπόλοιπες περιπτώσεις η παραπάνω συμπεριφορά δεν παρατηρείται στον ίδιο (έντονο) βαθμό. Ο λόγος που συμβαίνει αυτό είναι ότι έχει χρησιμοποιηθεί η εκθετική συνάρτηση για τη μείωση του ρυθμού μάθησης, οπότε οι τιμές του ρυθμού εκμάθησης μειώνονται πολύ ταχύτερα, σε αντίθεση με το δεύτερο τεχνητό dataset όπου χρησιμοποιήθηκε η γραμμική συνάρτηση.

Το μέτρο της παραμόρφωσης (distortion measure) αναμένουμε επίσης να φθίνει κατά την εκπαίδευση του χάρτη. Τα γραφήματα που λαμβάνουμε δείχνουν πως και στην περίπτωση αυτού του μέτρου παρατηρούνται πολύ θετικά αποτελέσματα. Φαίνεται συνεπώς, πως το γεγονός ότι ο Self Organizing Hidden Markov Map έλκει την καταγωγή του από το SOM συμβάλλει κατά έναν τρόπο στο να περιγράφεται κατά την εκπαίδευσή του ικανοποιητικά καλά από το μέτρο παραμόρφωσης. Κάποιες ταλαντώσεις που παρατηρούνται στο αρχικό στάδιο της εκπαίδευσης, όπως στο πρώτο τεχνητό σύνολο δεδομένων (παρατηρήθηκαν και σε άλλα πειράματα που τελέσθηκαν), δεν αφορούν κακή συμπεριφορά του μέτρου αλλά εντάσσονται στην γενικότερα ασταθή συμπεριφορά του SOHMMM κατά τις πρώτες εποχές (μεγάλη τιμή του ρυθμού μάθησης και της συνάρτησης γειτονιάς).

Το σφάλμα κβαντισμού (quantization error) φθίνει επίσης κατά την εξέλιξη της εκπαίδευσης, όπως αναμένεται. Η παρόμοια, σχετικά, μορφή του με αυτή του μέτρου παραμόρφωσης, αντικατοπτρίζεται στην όμοια συμπεριφορά τους σε όλες τις περιπτώσεις εκπαίδευσης, με παρουσία τοπικών ακρότατων στα ίδια σημεία. Θα πρέπει σε αυτό το σημείο να σημειωθεί πως το σφάλμα κβαντισμού είναι προφανώς (βλ. Κεφάλαιο 5) κατά πολύ ταχύτερο στον υπολογισμό του από το μέτρο παραμόρφωσης, οπότε δοθείσης της κοινής συμπεριφοράς τους, είναι σαφώς προτιμότερο ανάμεσα στα δύο. Τέλος, τονίζεται ότι η θετικότατη συμπεριφορά του σφάλματος κβαντισμού παρατηρήθηκε σε όλα τα πειράματα που διεξήχθησαν (ακόμα και σε αυτά που τελικά δεν παρουσιάζονται) κάτι που οφείλεται στο γεγονός ότι το SOHMMM αποτελεί έναν ακόμη αλγόριθμο κβαντισμού του χώρου εισόδου (ο όρος κβαντισμός διανυσμάτων προφανώς αποφεύγεται αφού ο χώρος εισόδου του SOHMMM αποτελείται από συμβολοακολουθίες και όχι διανύσματα).

Το επόμενο μέτρο της ανάλυσης είναι τοπογραφικό σφάλμα. Σημειώνεται καταρχήν, πως για το μέτρο αυτό δεν αναμέναμε αρχικά να υπάρχει κάποια μονοτονία αφού μετρά τις παραβιάσεις της γειτονιάς (βλέπε Κεφάλαιο 5) και συγκεκριμένα το αν η πρώτη και η δεύτερη BMU κάθε ακολουθίας είναι γειτονικές. Έτσι, αρχικά το εισάγαμε στην έρευνά μας για τη μονοτονία εντελώς διερευνητικά. Επίσης σημειώνεται ότι το απλό αυτό μέτρο είναι εξαιρετικά ευαίσθητο και δεν ενδιαφέρεται για το αν μια παραβίαση σημαίνει πως οι δύο BMU απέχουν μεταξύ τους παραδείγματος χάριν έναν κόμβο ή όλοι τη διάσταση του χάρτη. Παρόλ' αυτά, και στην περίπτωση αυτού του μέτρου τα ληφθέντα αποτελέσματα είναι πάρα πολύ θετικά. Στην περίπτωση των τεχνητών συνόλων δεδομένων παρατηρούνται μεγάλες

ταλαντώσεις που για την περίπτωση του δευτέρου συνόλου καταλήγουν να οδηγούν σε μια μη μονότονη συμπεριφορά του μέτρου. Πάντως η επίδοση του τοπογραφικού σφάλματος για τα υπόλοιπα datasets (ιδίως για τα πραγματικά) καθιστά αρκετά επιθυμητή τη χρήση του. Ιδιαίτερα, πρέπει να τονιστεί ότι από τα 8 μέτρα οργάνωσης που χρησιμοποιήθηκαν, μόνο αυτό είναι σε θέση να μας δώσει με σαφή τρόπο μια ποσοτική εκτίμηση της απόκλισης της χαρτογράφησης από την τελειότητα. Αυτό συμβαίνει εφόσον γνωρίζουμε ότι μια τέλεια χαρτογράφηση θα είχε τοπογραφικό σφάλμα 0. Για τα υπόλοιπα 7 μέτρα τέτοιο ανώτατο (ή κατώτατο) όριο δεν υφίσταται ή τουλάχιστον δεν είναι εκ των προτέρων γνωστή η τιμή του (μόνο για την περίπτωση του σφάλματος κβαντισμού θα μπορούσαμε να πούμε πως ιδανικά αναμένουμε να καταλήξει στην τιμή 0 κάτι που όμως δεν συμβαίνει στην πραγματικότητα για ποικίλους λόγους, κυρίως επειδή ο αλγόριθμος του SOHMMM δεν είναι ένας "καθαρόαιμος" αλγόριθμος κβαντισμού του χώρου εισόδου).

Ένα απλό μέτρο που εισήχθη για πρώτη φορά στη μελέτη μας και που από όσο γνωρίζουμε δεν έχει χρησιμοποιηθεί προηγουμένως, αφορά των αριθμό των ακολουθιών που αλλάζουν κόμβο από μια εποχή σε μια άλλη. Το μέτρο αυτό, την εξέλιξη του οποίου επιλέξαμε να απεικονίσουμε με ραβδόγραμμα, αποτελεί εν τέλει έναν ακόμα ικανοποιητικό δείκτη της διαδικασίας οργάνωσης, μειούμενο με την εξέλιξη της εκπαίδευσης. Παρατηρούμε πως σε όλες τις περιπτώσεις το μέτρο αυτό πλησιάζει την τιμή 0 κατά το τέλος της εκπαίδευσης. Η τιμή αυτή (μηδέν) ερμηνεύεται ως ένας τερματισμός της διαδικασίας εκπαίδευσης αφού πλέον τα δεδομένα δεν αλλάζουν κόμβο. Σημειώνεται πως στην περίπτωση του πρώτου πραγματικού συνόλου δεδομένων αλλά και σε άλλες, ειδικές περιπτώσεις εκπαίδευσης που μελετήθηκαν και δεν παρουσιάζονται και για τις οποίες τα δεδομένα είχαν πολύ σαφή διαχωρισμό μεταξύ τους η τιμή αυτή έγινε δυνατόν να επιτευχθεί ακόμη και πριν το τέλος της εκπαίδευσης. Στην περίπτωση του πρώτου πραγματικού συνόλου δεδομένων, εντούτοις, παρατηρούμε ότι ενώ η διάταξη έχει επιτευχθεί ικανοποιητικά (βλέπε εικόνα 6.1) το μέτρο αυτό διατηρεί πολύ υψηλή τιμή. Το παραπάνω συμβαίνει για τον εξής λόγο. Στο πρώτο dataset το πλέγμα που χρησιμοποιήθηκε είναι 81 κόμβων ενώ ο αριθμός των ακολουθιών που χρησιμοποιήθηκαν είναι 150, συνεπώς το πλέγμα είναι αρκετά μεγάλο (>1/2 #δεδομένων). Παράλληλα, τα δεδομένα οργανώνονται σε τρεις συστάδες που είναι πολύ ανόμοιες μεταξύ τους. Αυτό που συμβαίνει εντέλει, είναι ότι ο χάρτης περιέχει μεγάλες περιοχές προς περιγραφή της κάθε συστάδας με τα HMM κάθε συστάδας να έχουν πολύ όμοιες τιμές παραμέτρων. Έτσι οι ακολουθίες αλλάζουν διαρκώς (σε κάθε εποχή) κόμβο παραμένοντας πάντως στην ίδια συστάδα. Τα παραπάνω επιβεβαιώθηκαν από την εποπτεία των συγκεκριμένων κόμβων που άλλαξαν ορισμένες ακολουθίες καθώς και από τη χρήση μικρότερων πλεγμάτων.

Τέλος, αναφέρονται τα συμπέρασμα από τη χρήση της εντροπίας ενεργοποίησης νευρώνα. Το μέγεθος αυτό αναμένουμε να αυξάνεται με την εξέλιξη της εκπαίδευσης. Η απεικόνιση έγινε σε λογαριθμική κλίμακα επειδή η εντροπία αυξάνονταν κατά τη διάρκεια της διαδικασίας διάταξης του χάρτη κατά αρκετές δεκάδες τάξεις μεγέθους. Η συμπεριφορά του μέτρου αυτού είναι σύμφωνη με την επιθυμητή σε όλες τις περιπτώσεις δείχνοντας έτσι με χαρακτηριστικό τρόπο πώς αυξάνεται η πιθανότητα ενεργοποίησης των νευρώνων από το σύνολο δεδομένων εισόδου όσο εξελίσσεται η εκπαίδευση. Με άλλα λόγια η εντροπία καταφέρνει να δείξει ότι το πλέγμα του SOHMMM θα "εξαπλωθεί" στο χώρο εισόδου με τρόπο που να τον περιγράψει ικανοποιητικά και κάθε HMM θα περιγράφει κάποιο "τμήμα" του.

Ένα τελικό συνολικό σχόλιο που μπορεί να γίνει μετά την ανάλυση που ακολούθησε είναι πως τα 8 μελετηθέντα ως προς τη μονοτονία μέτρα μπορούν εν γένει να χωριστούν ως προς την πληροφορία που μας δίνουν κατά τη διάρκεια της εκπαίδευσης στις εξής κατηγορίες:

- ποιότητα της απεικόνισης από τοπογραφικής απόψεως (παραβιάσεις γειτονιάς, ομαλότητα απεικόνισης,κλπ). Σε αυτή την κατηγορία ανήκουν τα: C-measure, Demartines-Herault, distortion measure, τοπογραφικό σφάλμα. Προτιμητέο θα μπορούσαμε να πούμε πως είναι το τοπογραφικό σφάλμα, τόσο λόγω της πολύ καλής επίδοσής του όσο και εξαιτίας της εξαιρετική απλότητας και ταχύτητας υπολογισμού του.
- ύπαρξη συμπαγών συστάδων (κόμβων) και μεγάλη ανομοιότητα μεταξύ των συστάδων (κόμβων). Σε αυτή την κατηγορία ανήκουν το σφάλμα κβαντισμού και το κριτήριο Calinski-Harabasz. Πιο ταχύ στον υπολογισμό καθώς και πιο σταθερά βέλτιστο στην επίδοσή του είναι το σφάλμα κβαντισμού, πάντως το Calinski Harabasz μετρά και την ανομοιότητα μεταξύ των κόμβων (between group sum of squares).
- εξέλιξη και ολοκλήρωση της διαδικασίας εκπαίδευσης. Ο αριθμός των ακολουθιών που αλλάζουν κόμβο σε κάθε εποχή ανήκει σε αυτή την κατηγορία.
- αύξηση της πιθανότητας ενεργοποίησης του κάθε νευρώνα από το σύνολο των δεδομένων εισόδου. Στην κατηγορία αυτή ανήκει η εντροπία ενεργοποίησης νευρώνα.

Συνεπώς, ανάλογα με τις πληροφορίες που θέλουμε να αποκομίσουμε για την εκπαίδευση καθώς και τους υπολογιστικούς πόρους μπορούμε να επιλέξουμε ένα μέτρο από τα παραπάνω ή έναν συνδυασμό τους.

6.3.2 Μελέτη ως προς των αριθμό των εποχών

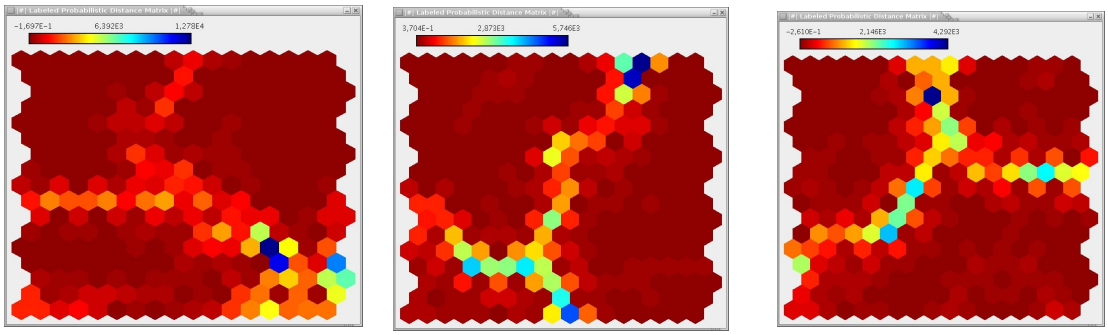
Σε αυτή την παράγραφο καταγράφονται και αναλύονται τα αποτελέσματα που ελήφθησαν από τις προσομοιώσεις ως προς τον αριθμό των εποχών. Σημειώνεται σε αυτό το σημείο ότι λόγω του γεγονότος ότι η διαδικασία της εκπαίδευσης ενός SOHMMM είναι μια πιθανοτική διαδικασία, τα αποτελέσματα εκπαιδεύσεων με ίδιες παραμέτρους μπορεί να είναι λίγο διαφορετικά (χωρίς μεγάλες πάντως αποκλίσεις). Η μεθοδολογία που ακολουθήθηκε περιελάμβανε την εκτέλεση 5 πειραμάτων για κάθε εκπαίδευση διαφορετικών παραμέτρων και τη λήψη μετέπειτα του μέσου όρου αυτών ώστε να είναι πιο αντιπροσωπευτικές οι τιμές. Κάθε πείραμα εκτελέστηκε με τις παραμέτρους που εκτελέστηκε και για την περίπτωση της μελέτης της μονοτονίας μεταβάλλοντας κάθε φορά μόνο την παράμετρο που μας ενδιέφερε να μελετήσουμε (πχ για το πρώτο dataset των globin στη μελέτη ως προς εποχές το πλέγμα έμεινε σταθερό 7x4 και ο αριθμός HMM ίσος με 11, αλλάζοντας κάθε φορά μόνο τον αριθμό εποχών). Επίσης, οι ληφθείσες τιμές έχουν ακρίβεια 8 δεκαδικών ψηφίων. Τα ανωτέρω ισχύουν και για τις περιπτώσεις μελέτης ως προς τα HMM και τη διάσταση πλέγματος.

Στις προσομοιώσεις έγινε χρήση όλων των μέτρων εκτός του αριθμού των ακολουθιών που άλλαξαν κόμβο και του τοπογραφικού γινομένου λόγω του ότι κρίθηκαν ακατάλληλα για την ποσοτικοποίηση του βέλτιστου αριθμού εποχών εκπαίδευσης. Τα αποτελέσματα ελήφθησαν για έναν ενδεικτικό αριθμό εποχών. Αυτό που επιθυμούμε να παρατηρήσουμε είναι για ποιο αριθμό εποχών το κάθε κριτήριο επιτυγχάνει βέλτιστη τιμή και αν το αποτέλεσμα αυτό συνάδει με την ποιότητα διάταξης του χάρτη. Επίσης, εξετάζεται αν υπάρχει συμφωνία μεταξύ των κριτηρίων στα αποτελέσματά τους καθώς και αν υπάρχει εύλογη ερμηνεία των ληφθέντων αποτελεσμάτων. Ένα γεγονός που αξίζει να αναφερθεί είναι ότι στις περισσότερες περιπτώσεις αύξηση του αριθμού εποχών εκπαίδευσης προκαλεί λιγότερο ή περισσότερο καλύτερη διάταξη του χάρτη. Σε κάποιες όμως περιπτώσεις, παρατηρείται υπερ-εκπαίδευση του χάρτη και χειροτέρευση της διάταξης. Αυτό που εν τέλει αναζητούμε είναι αν τα μέτρα οργάνωσης που αναπτύχθηκαν μπορούν να περιγράψουν αυτά τα φαινόμενα και να μας δώσουν μια ποσοτική πληροφορία περί αυτών.

6.3.2.1 Τεχνητά σύνολα δεδομένων

Πρώτο τεχνητό σύνολο δεδομένων (150 δείγματα)

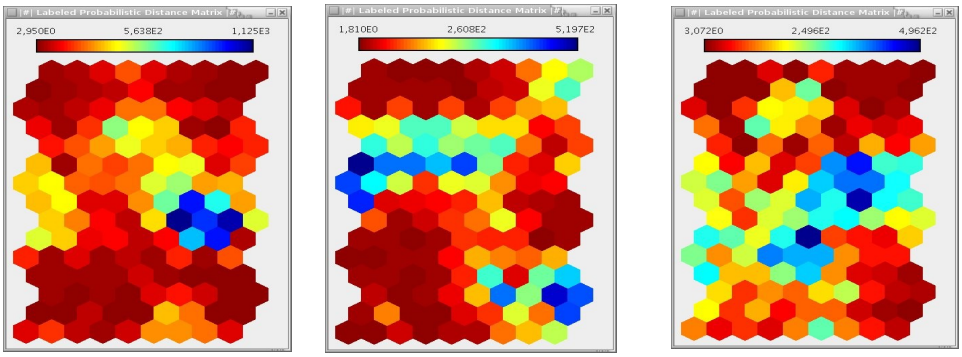
Εποχές	10	20	40	100	1000
σφάλμα κβαντισμού	8,50E-01	1,47E+00	1,38E+00	1,15E+00	8,50E-01
εντροπία	6,79E-39	1,07E-40	1,45E-38	9,82E-39	1,89E-38
τοπογραφικό σφάλμα	4,30E-01	4,30E-01	3,07E-01	3,20E-01	3,53E-01
παραμόρφωση	4,17E+00	6,64E+00	6,44E+00	5,18E+00	3,76E+00
C-measure	1,83E+08	2,84E+08	2,74E+08	2,79E+08	1,34E+08
Calinski-Harabasz	1,75E+07	2,8847611E7	2,93E+07	3,04E+07	1,33E+07
Demartines-Herault	1,14E+02	1,08E+02	1,04E+02	1,06E+02	1,09E+02



α. 10 εποχές *β.20 εποχές* *γ.1000 εποχές*
Εικόνα 6.11 Για το πρώτο τεχνητό dataset, δεν παρατηρούνται φαινόμενα overtraining

Δεύτερο τεχνητό σύνολο δεδομένων (40 δείγματα)

Εποχές	5	10	20	100	1000
σφάλμα κβαντισμού	3,30E+00	2,56E+00	3,24E+00	2,63E+00	2,62E+00
εντροπία	2,69E-33	6,19E-32	5,37E-32	2,26E-31	1,67E-31
τοπογραφικό σφάλμα	7,50E-02	1,00E-01	1,25E-01	2,00E-01	1,75E-01
παραμόρφωση	1,26E+01	1,31E+01	1,19E+01	1,05E+01	9,92E+00
C-measure	1,27E+07	2,01E+07	1,97E+07	2,21E+07	2,29E+07
Calinski-Harabasz	7,02E+04	1,24E+05	1,18E+05	1,30E+05	1,30E+05
Demartines-Herault	4,81E+01	4,33E+01	4,60E+01	4,80E+01	4,91E+01



α. 5 εποχές *β.10 εποχές* *γ.20 εποχές*

Εικόνα 6.12 Φαινόμενο overtraining για το δεύτερο τεχνητό dataset

Τρίτο τεχνητό σύνολο δεδομένων (30 δείγματα)

Εποχές	10	20	40	100	1000
σφάλμα κβαντισμού	5,94E+00	5,91E+00	5,88E+00	5,92E+00	5,84E+00
εντροπία	5,27E-70	1,41E-67	2,01E-66	2,39E-71	6,96E-65
τοπογραφικό σφάλμα	7,67E-01	5,33E-01	3,00E-01	3,33E-01	4,00E-01
παραμόρφωση	2,81E+01	2,65E+01	2,56E+01	2,41E+01	2,72E+01
C-measure	4,10E+06	4,29E+06	3,55E+06	5,51E+06	4,39E+06
Calinski-Harabasz	2,11E+04	2,23E+04	1,74E+04	2,50E+04	2,20E+04
Demartines-Herault	2,73E+01	2,39E+01	2,62E+01	2,77E+01	2,86E+01

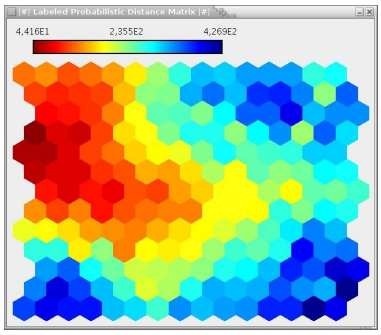
6.3.2.2 Globins

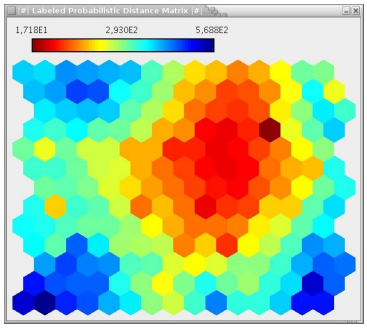
Globins 1 (30 δείγματα)

Εποχές	5	10	20	100	1000
σφάλμα κβαντισμού	1,28E+01	6,28E+00	1,30E+01	1,16E+01	1,51E+01
εντροπία	2,37E-100	1,42E-94	1,32E-92	5,55E-91	5,65E-82
τοπογραφικό σφάλμα	1,67E-01	5,67E-02	1,67E-01	6,67E-02	1,67E-01
παραμόρφωση	5,44E+01	2,08E+01	4,71E+01	4,20E+01	5,70E+01
C-measure	1,34E+06	2,21E+06	2,85E+06	3,85E+06	3,94E+06
Calinski-Harabasz	7,69E+03	1,30E+04	1,75E+04	2,42E+04	2,45E+04
Demartines-Herault	1,72E+01	1,44E+01	1,44E+01	1,33E+01	1,38E+01

Globins 2 (560 δείγματα)

Εποχές	10	20	40	100	1000
σφάλμα κβαντισμού	2,49E+00	2,50E+00	2,47E+00	2,46E+00	2,45E+00
εντροπία	5,83E-122	2,95E-124	2,53E-122	6,42E-121	5,73E-119
τοπογραφικό σφάλμα	1,25E-01	2,57E-01	2,25E-01	1,23E-01	1,48E-01
παραμόρφωση	1,02E+01	1,08E+01	1,00E+01	1,02E+01	1,03E+01
C-measure	1,67E+07	1,79E+07	1,82E+07	2,06E+07	2,10E+07
Calinski-Harabasz	1,50E+06	1,62E+06	1,67E+06	1,84E+06	1,89E+06
Demartines-Herault	4,05E+01	4,09E+01	4,04E+01	4,28E+01	4,27E+01





α. 10 εποχές *β.1000 εποχές (560000 βήματα)*

Εικόνα 6.13 Η οπτικοποίηση επιβεβαιώνει τα μέτρα οργάνωσης, ότι δηλαδή αύξηση των εποχών εκπαίδευσης δεν βελτιώνει η χειροτερεύει τη διάταξη του χάρτη.

6.3.2.3 Συμπεράσματα

Παρατηρώντας τα αποτελέσματα των προσομοιώσεων σε συνάρτηση με τις αντίστοιχες διατάξεις που επιτυγχάνει ο χάρτης μπορούμε να καταλήξουμε σε κάποια συμπεράσματα ως προς τη δυνατότητα των μελετηθέντων κριτηρίων να μας βοηθήσουν να αποφανθούμε σχετικά με τον βέλτιστο αριθμό εποχών για την εκπαίδευση του χάρτη. Εν γένει, τα αποτελέσματα αυτά φανερώνουν ένα πολύ μεγάλο πλήθος πληροφοριών ως προς φαινόμενα overtraining ή επιλογής του αριθμού εποχών και θα προσπαθήσουμε ενδεικτικά να παρουσιάσουμε τα πιο ενδιαφέροντα συμπεράσματα.

Όσον αφορά το σφάλμα κβαντισμού, παρατηρούμε ότι επιτυγχάνει με μικρές αποκλίσεις να βρει τον βέλτιστο αριθμό εποχών εκπαίδευσης. Ενδιαφέρον παρουσιάζει το πρώτο τεχνητό σύνολο δεδομένων όπου παρατηρούμε ότι το σφάλμα κβαντισμού για περίοδο εκπαίδευσης 10 εποχών επιτυγχάνει μια τιμή που το δίκτυο δείχνει να είναι ικανό να ξανα-αποκτήσει μόνο για πολύ μεγάλο αριθμό εποχών (~1000). Το άμεσο συμπέρασμα από την παραπάνω παρατήρηση είνΙαι η προτίμηση εκπαίδευσης για αριθμό 10 εποχών. Μια ακόμη ενδιαφέρουσα παρατήρηση λαμβάνεται από το πρώτο σύνολο δεδομένων των globins όπου παρατηρείται το φαινόμενο του overtraining για τον χάρτη. Αρχικά, παρατηρείται στις 10 έποχες ελάχιστο, που αντικατοπτρίζει την δυνατότητα γρήγορης διάταξης των δεδομένων

από το SOHMMM ενώ στη συνέχεια η τιμή του σφάλματος κβαντισμού αυξομειώνεται καταλήγοντας σε τοπικό ελάχιστο στις 100 εποχές για να αυξηθεί ξανά υπέρμετρα για 1000 εποχές δείχνοντας έτσι ότι ο χάρτης "υπερ-εκπαιδεύεται".

Η εντροπία αντίθετα, επιδεικνύει μια όχι και τόσο επιθυμητή συμπεριφορά μειούμενη με την αύξηση του αριθμού των εποχών. Το γεγονός αυτό παρατηρείται, με μικρές αποκλίσεις, για όλα τα σύνολα δεδομένων. Η εξέλιξη αυτή θα μπορούσε ίσως να είναι αναμενόμενη από το γεγονός ότι η εντροπία περιγράφει το πόσο μεγάλη είναι η μέση πιθανότητα ενεργοποίησης των νευρώνων. Συνεπώς, είναι ίσως λογικό να αναμένουμε η εντροπία να μην είναι σε θέση να περιγράψει φαινόμενα overtraining αφού με την αύξηση του αριθμού εποχών, η μέση πιθανότητα ενεργοποίησης ενός νευρώνα αναμένεται να αυξάνεται.

Το τοπογραφικό σφάλμα, αν και όχι εντελώς ικανοποιητικά, επιτυγχάνει να μας δώσει κάποια πληροφορία για τον προτιμώμενο αριθμό εποχών. Στην περίπτωση του πρώτου συνόλου δεδομένων των globins τα αποτελέσματά του συμφωνούν με αυτά του σφάλματος κβαντισμού. Εν γένει πάντως, δείχνει να μην έχει τη δυνατότητα περιγραφής φαινομένων overtraining, μειούμενο σε γενικές γραμμές με την αύξηση του αριθμού εποχών εκπαίδευσης.

Το μέτρο παραμόρφωσης δείχνει μια παρόμοια συμπεριφορά με αυτή του σφάλματος κβαντισμού, φανερώνοντας έτσι την ίδια πληροφορία όσον αφορά τις διατάξεις των χαρτών. Στις περιπτώσεις του πρώτου τεχνητού συνόλου δεδομένων και του πρώτου συνόλου πρωτεϊνών παρατηρείται πανομοιότυπη συμπεριφορά, όπως και στην περίπτωση του δευτέρου συνόλου πρωτεϊνών όπου η ομοιότητα των ληφθείσων τιμών για διαφορετικές εποχές συστήνει ουσιαστικά ότι μπορούμε να χρησιμοποιήσουμε λίγες εποχές πετυχαίνοντας ικανοποιητικά αποτελέσματα διάταξης του χάρτη (σημειώνουμε ότι το ίδιο συμπέρασμα προκύπτει και από το σφάλμα κβαντισμού).

Το μέτρο C καθώς και το κριτήριο Calinski-Harabasz δεν επιδεικνύουν σε γενικές γραμμές επιθυμητή συμπεριφορά με τις αυξομοιώσεις των τιμών τους να μη ακολουθούν τις οπτικοποιήσεις ή τα άλλα μέτρα οργάνωσης. Τα μέτρα αυτά λοιπόν, τουλάχιστον με βάση τα αποτελέσματα που λήφθηκαν, δεν συστήνονται για χρήση σχετικά με την απόφανση περί του βέλτιστου αριθμού εποχών εκπαίδευσης.

Το μέτρο των Demartines-Herault έρχεται σε συμφωνία με τα ποιοτικά αποτελέσματα για το δεύτερο τεχνητό και το δεύτερο πρωτεϊνικό σύνολο δεδομένων αλλά δεν έχει επιθυμητή συμπεριφορά στις άλλες περιπτώσεις.

Εν τέλει, θα μπορούσαμε να πούμε, με βάση τα μέτρα οργάνωσης που μελετήθηκαν, ότι στην απόφαση σχετικά με τον βέλτιστο αριθμό εποχών εκπαίδευσης θα μπορούσαμε να λάβουμε υπόψιν ως δείκτες το σφάλμα κβαντισμού και το μέτρο της παραμόρφωσης, κατά περίπτωση δε και τα τοπογραφικό σφάλμα. και μέτρο Demartines-Herault. Πάντως, σίγουρα τα μέτρα αυτά δεν μπορούν να μας δώσουν μια αποδεδειγμένα σωστή απάντηση, όπως προέκυψε και από τα παραπάνω και η πληροφορία που συνάγουμε θα πρέπει να μελετάται προσεκτικά, σε συνάρτηση και με άλλους παράγοντες (πχ οπτικοποίηση του SOHMMM).

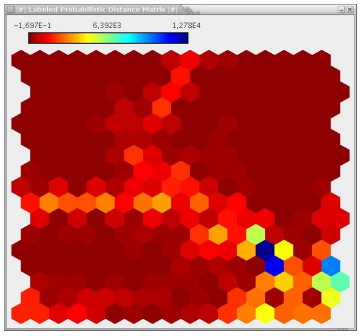
6.3.3 Μελέτη ως προς των αριθμό των κρυμμένων καταστάσεων των HMMs

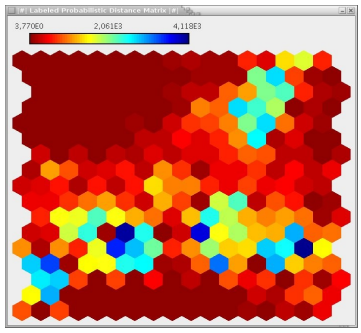
Τα πειράματα σχετικά με την επιλογή του βέλτιστου αριθμού καταστάσεων για τα μαρκοβιανά μοντέλα, πραγματοποιήθηκαν με μεθοδολογία όμοια με προηγουμένως. Μελετώνται κάποιες ενδεικτικές περιπτώσεις, λαμβάνεται μέσος όρος των παρατηρούμενων αποτελεσμάτων (πέντε πειράματα για κάθε περίπτωση) και επιχειρείται να διερευνηθεί η δυνατότητα των μέτρων οργάνωσης να συμβάλλουν στην απόφαση σχετικά με αυτή την παράμετρο καθώς και να ερμηνευθούν τα ληφθέντα αποτελέσματα.

6.3.3.1 Τεχνητά σύνολα δεδομένων

Πρώτο τεχνητό σύνολο δεδομένων (150 δείγματα)

#HMMs	1	3	5	10
σφάλμα κβαντισμού	8,50E-01	8,48E-01	8,46E-01	8,40E-01
εντροπία	6,79E-39	3,46E-39	7,25E-39	4,59E-39
τοπογραφικό σφάλμα	4,27E-01	5,80E-01	5,13E-01	1,93E-01
παραμόρφωση	4,17E+00	4,06E+00	3,99E+00	3,56E+00
C-measure	1,83E+08	1,54E+08	1,59E+08	1,23E+08
Calinski-Harabasz	1,75E+07	1,63E+07	1,79E+07	1,31E+07
Demartines-Herault	1,14E+02	1,06E+02	9,91E+01	1,03E+02





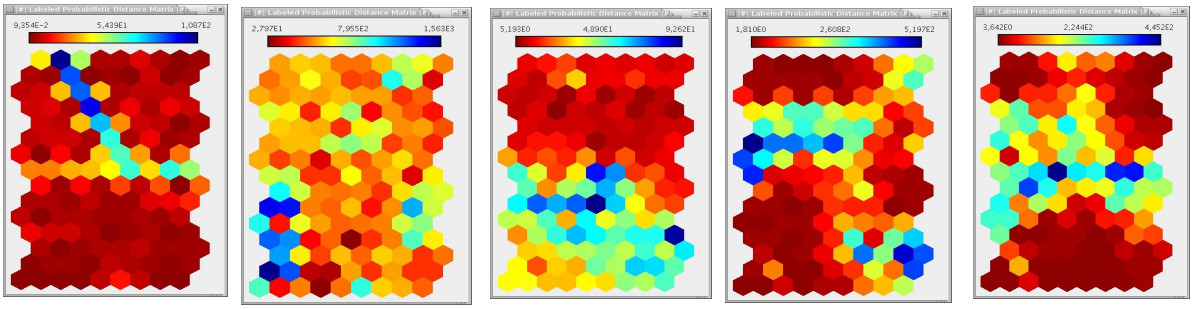
α. 1 κατάστασης *β. 10 καταστάσεων*

Εικόνα 6.14 SOHMMMs με HMM διαφορετικών καταστάσεων για το πρώτο τεχνητό σύνολο δεδομένων

Δεύτερο τεχνητό σύνολο δεδομένων (40 δείγματα)

#HMMs	1	5	10	12
σφάλμα κβαντισμού	5,53E+00	1,27E+01	2,56E+00	2,81E+00
εντροπία	1,46E-75	1,13E-73	6,19E-72	1,15E-70
τοπογραφικό σφάλμα	2,25E-01	5,25E-01	1,50E-01	7,50E-02
παραμόρφωση	2,21E+01	7,55E+01	1,40E+01	1,00E+01
C-measure	2,41E+06	1,91E+07	1,92E+07	1,91E+07
Calinski-Harabasz	5,35E+03	1,07E+05	1,11E+04	1,05E+05
Demartines-	4,50E+01	3,95E+01	3,70E+01	4,59E+01

Herault				



α. 1 κατάστασης β.3 καταστάσεων γ. 5 καταστάσεων δ. 10 καταστάσεων ε. 12 καταστάσεων
Εικόνα 6.15 SOHMMMs με HMM διαφορετικών καταστάσεων για το δεύτερο τεχνητό σύνολο δεδομένων.

Τρίτο τεχνητό σύνολο δεδομένων (30 δείγματα)

#HMMs	1	3	5	10
σφάλμα κβαντισμού	7,96E+00	6,39E+00	5,88E+00	5,82E+00
εντροπία	1,78E-137	7,79E-97	2,01E-66	7,12E-66
τοπογραφικό σφάλμα	6,00E-01	7,33E-01	3,00E-01	3,67E-01
παραμόρφωση	3,97E+01	2,99E+01	2,56E+01	2,34E+01
C-measure	1,62E+07	5,02E+06	3,55E+06	5,58E+06
Calinski-Harabasz	5,75E+04	2,44E+04	1,74E+04	2,71E+04
Demartines-Herault	2,63E+01	2,50E+01	2,62E+01	2,76E+01

6.3.3.2 Πραγματικά σύνολα δεδομένων

Globins 1 (30 δείγματα)

#HMMs	1	3	5	11	22
σφάλμα κβαντισμού	1,03E+01	2,44E+01	1,78E+01	6,28E+00	4,69E+00
εντροπία	3,73E-124	3,67E-119	1,35E-106	1,42E-94	1,31E-115
τοπογραφικό σφάλμα	1,00E-01	3,67E-01	3,00E-01	6,00E-02	3,67E-01
παραμόρφωση	3,88E+01	5,50E+01	6,45E+01	2,08E+01	1,91E+01
C-measure	3,85E+05	7,41E+05	1,41E+06	2,27E+06	2,60E+06

Calinski-Harabasz	1,62E+03	3,59E+03	7,23E+03	1,32E+04	1,74E+04
Demartines-Herault	2,60E+01	1,99E+01	1,79E+01	1,52E+01	1,55E+01

Globins 2 (560 δείγματα)

#HMMs	1	3	5	11	22
σφάλμα κβαντισμού	2,85E+00	3,00E+00	2,85E+00	2,49E+00	2,17E+00
εντροπία	8,85E-145	9,73E-155	7,82E-145	5,83E-122	2,25E-104
τοπογραφικό σφάλμα	2,25E-01	3,43E-01	2,80E-01	1,25E-01	1,00E-01
παραμόρφωση	1,18E+01	1,31E+01	1,12E+01	1,03E+01	8,69E+00
C-measure	2,10E+06	5,40E+06	9,60E+06	1,67E+07	2,45E+07
Calinski-Harabasz	1,30E+05	3,98E+05	7,56E+05	1,50E+06	2,53E+06
Demartines-Herault	6,92E+01	4,86E+01	6,04E+01	4,05E+01	4,08E+01

6.3.3.3 Συμπεράσματα

Στα πειράματα που πραγματοποιήθηκαν ως προς αυτή την παράμετρο έγινε δυνατόν να εξαχθούν σαφή αποτελέσματα σχετικά με τις δυνατότητες των μέτρων οργάνωσης. Για να αποφανθούμε σχετικά με το επιτυχές ήταν ένα μέτρο οργάνωσης και πόσο έγκυρες είναι οι ενδείξεις που παράγει, στηριχθήκαμε τόσο στην πρότερη γνώση της κατανομής των δεδομένων (για τα τεχνητά δεδομένα) όσο και στη μορφή του χάρτη μετά το τέλος της εκπαίδευσης (με τη βοήθεια του Probabilistic Distance Matrix) και τη συνάφεια των αποτελεσμάτων μεταξύ των διαφορετικών μέτρων.

Το σφάλμα κβαντισμού επιδεικνύει μια πάρα πολύ θετική συμπεριφορά αποτυπώνοντας σε όλες τις περιπτώσεις τον βέλτιστο αριθμό καταστάσεων. Ιδιαίτερα σχετικά με τα πραγματικά σύνολα δεδομένων, παρουσιάζουν ενδιαφέρον οι ληφθείσες τιμές. Παρατηρούμε ότι βέλτιστη τιμή για το δεύτερο dataset επιτυγχάνεται για 22 καταστάσεις (όσα και τα γράμματα του αλφαβήτου για τις πρωτεΐνες). Το παραπάνω ίσως δείχνει πως ο χάρτης διατάσσεται καλύτερα-γρηγορότερα αν επιλέξουμε HMMs με αριθμό κρυμμένων καταστάσεων κοντά στο μέγεθος του αλφάβητου. Πάντως, και με μαρκοβιανά μοντέλα 11 καταστάσεων η διάταξη είναι αρκετά ικανοποιητική με την τιμή του σφάλματος κβαντισμού να μην είναι ιδιαίτερα χειρότερη.

Η εντροπία καταφέρνει να περιγράψει με εύλογο τρόπο τα αποτελέσματα σχετικά με τις πρωτεΐνες αλλά δεν έχει ικανοποιητική απόδοση στα τεχνητά σύνολα δεδομένων. Το μέγεθος αυτό συστήνεται συνεπώς προς χρήση για την απόφανση ως προς τον αριθμό καταστάσεων των HMM με κάποια επιφύλαξη.

Το τοπογραφικό σφάλμα έχει εν γένει ικανοποιητική απόδοση, παρόμοια με αυτή του σφάλματος κβαντισμού. Ιδιαίτερο ενδιαφέρον παρουσιάζει η ομοιότητα των δύο κριτηρίων ως προς τη συμπεριφορά, παραδείγματος χάριν για τα πρωτεϊνικά σύνολα δεδομένων. Ένα σημείο που αξίζει αναφοράς είναι η περίπτωση του πρώτου τεχνητού συνόλου δεδομένων όπου τόσο το σφάλμα κβαντισμού όσο και το τοπογραφικό σφάλμα (η παραμόρφωση επίσης)

παρουσιάζουν την ελάχιστη τιμή τους για HMM δέκα καταστάσεων. Η οπτικοποίηση μέσω του PDM έδειξε να συμφωνεί με το συμπέρασμα αυτό, αφού η διάταξη που επιτεύχθηκε ήταν καλύτερη για τις υπόλοιπες περιπτώσεις. Έτσι, αν και το dataset αυτό δημιουργήθηκε μέσω τριών απλών κατανομών (HMM μιας κατάστασης), εν τέλει φαίνεται πως ένα δίκτυο με HMM δέκα καταστάσεων είναι σε θέση να το περιγράψει καλύτερα. Φαίνεται συνεπώς, πως τα μέτρα οργάνωσης μπορούν σε κάποιες περιπτώσεις να δώσουν μια πολύτιμη πληροφορία σχετικά με την παραμετροποίηση του SOHMMM.

Για το μέτρο παραμόρφωσης τα συμπεράσματα σχετικά με την επίδοσή του είναι σχετικά παρόμοια με του τοπογραφικού σφάλματος. Πάντως και τα δύο μέτρα αποτυγχάνουν να βρουν τη σωστή βέλτιστη τιμή για το δεύτερο τεχνητό σύνολο δεδομένων (10) παρουσιάζοντας ελάχιστη τιμή για HMM δώδεκα καταστάσεων.

Το μέτρο C και το κριτήριο Calinski Harabasz έχουν ικανοποιητικές επιδόσεις για τα πραγματικά σύνολα δεδομένων (δείχνουν ότι οι 11 ή 22 καταστάσεις είναι προτιμητέες, με τις δεύτερες να υπερέχουν ελαφρώς) και για το δεύτερο τεχνητό σύνολο δεδομένων. Στα άλλα δύο τεχνητά σύνολα δεδομένων οι επιδόσεις δεν είναι οι επιθυμητές.

Το κριτήριο Demartines-Herault έχει επιδόσεις παρόμοιες με των δύο παραπάνω κριτηρίων. Δίνει μια πολύ ξεκάθαρη πληροφορία σχετικά με τα πραγματικά σύνολα δεδομένων και το δεύτερο τεχνητό dataset, αλλά στις άλλες δύο περιπτώσεις παρουσιάζει τιμές που δεν έρχονται σε συμφωνία με την ποιότητα της διάταξης

Καταλήγοντας, θα μπορούσαμε να θεωρήσουμε ως πιο αξιόπιστο το σφάλμα κβαντισμού με το τοπογραφικό σφάλμα και την παραμόρφωση να ακολουθούν. Τα μέτρα C, Calinski-Harabasz και Demartines-Herault θα ήταν καλό να λαμβάνονται υπόψιν, σε συνάρτηση πάντως με κάποια άλλα. Και σε αυτή την περίπτωση, θα μπορούσαμε να πούμε πως τα συμπεράσματα από τη χρήση αυτών των μέτρων μπορούν εν γένει να βοηθήσουν στην βέλτιστη παραμετροποίηση του SOHMMM, χωρίς όμως τα αποτελέσματά τους να ακολουθούνται τυφλά.

6.3.4 Μελέτη ως προς το μέγεθος του πλέγματος του SOHMMM

Στην ενότητα αυτή περιγράφονται τα αποτελέσματα των προσομοιώσεων που έγιναν σχετικά με την επιλογή βέλτιστου πλέγματος για το SOHMMM. Μελετήθηκαν τα εξής μέτρα οργάνωσης: σφάλμα κβαντισμού, εντροπία, τοπογραφικό σφάλμα, παραμόρφωση, τοπογραφικό γινόμενο. Τα υπόλοιπα μέτρα θεωρήθηκαν ακατάλληλα να διακρίνουν τη διαφορά μεταξύ διαφόρων πλεγμάτων (εξάλλου, τα C και Demartines-Herault δεν μπορούν να κανονικοποιηθούν ικανοποιητικά για διαφορετικά μεγέθη πλέγματος). Στόχος αυτής της διερεύνησης ήταν να βρεθούν τα μέτρα αυτά που μπορούν να διακρίνουν καλύτερα ποιο μέγεθος χάρτη είναι βέλτιστο, δοθέντος ενός συνόλου δεδομένων εισόδου.

Η απόφαση αυτή (περί του μεγέθους/διάστασης ενός χάρτη) δεν είναι πάντοτε εύκολη στην περίπτωση του SOHMMM, καθώς ένας πρόσθετος παράγοντας που υπεισέρχεται σε σχέση με το προγενέστερο SOM είναι ότι τα δεδομένα εισόδου είναι συμβολοακολουθίες και όχι διανύσματα και συνεπώς δεν έχουν μια σαφή αναφορά σε έναν χώρο. Με άλλα λόγια, το SOM πάντα καλούνταν να περιγράψει μια κατανομή διανυσμάτων εισόδου που προέρχονταν από έναν τυπικά ορισμένο χώρο (με την έννοια της τριγωνικής ανισότητας να υφίσταται). Στο τωρινό SOHMMM, τέτοιος χώρος δεν υφίσταται, τουλάχιστον τυπικά. Συνεπώς, δεν είναι πάντοτε ξεκάθαρο αν μια συστάδα που καταλήγει να καταλάβει μεγαλύτερο μέρος του χάρτη σε σχέση με μια άλλη, καταλαμβάνει και αντίστοιχα μεγαλύτερο χώρο από αυτή στον "χώρο" εισόδου. Τα παραπάνω καθιστούν την επιλογή πλέγματος μια σύνθετη εργασία, ακόμη και αν είναι γνωστή η κατανομή των δεδομένων εισόδου.

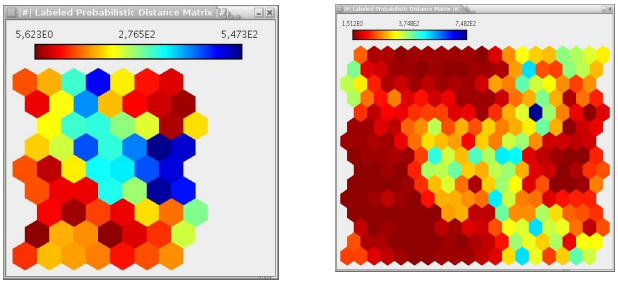
6.3.4.1 Τεχνητά σύνολα δεδομένων

Πρώτο τεχνητό σύνολο δεδομένων (150 δείγματα)

Πλέγμα	3x2	3x5	6x5	9x9	12x12	14x14
σφάλμα κβαντισμού	1,60E+01	8,77E+00	2,31E+00	8,50E-01	4,78E-01	3,51E-01
εντροπία	2,01E-43	9,09E-39	2,88E-39	6,79E-39	8,49E-39	9,22E-39
τοπογραφικό σφάλμα	6,70E-03	3,40E-01	4,40E-01	6,07E-01	5,40E-01	6,20E-01
παραμόρφωση	6,10E+01	4,17E+01	8,59E+00	4,17E+00	2,61E+00	2,26E+00
τοπογραφικό γινόμενο	-0,218	-0,201	-0,560	-0,770	-0,802	-0,595

Δεύτερο τεχνητό σύνολο δεδομένων (40 δείγματα)

Πλέγμα	3x2	4x4	4x5	5x8	10x8	10x12
σφάλμα κβαντισμού	1,71E+01	5,75E+00	6,50E+00	2,50E+00	1,93E+00	1,56E+00
εντροπία	1,55E-36	9,95E-33	9,83E-33	5,95E-30	1,79E-30	4,77E-31
τοπογραφικό σφάλμα	0,00E+00	1,25E-01	2,50E-01	7,50E-02	2,00E-01	3,50E-01
παραμόρφωση	5,45E+01	2,20E+01	2,25E+01	1,00E+01	7,84E+00	6,66E+00
τοπογραφικό γινόμενο	-0,0626	-0,278	-0,326	-0,426	-0,517	-0,489



Εικόνα 6.16 Πλέγματα 4x5 και 10x8 για το δεύτερο τεχνητό σύνολο δεδομένων

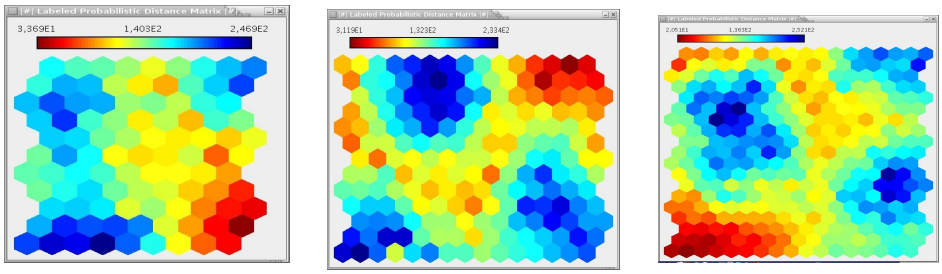
Τρίτο τεχνητό σύνολο δεδομένων (30 δείγματα)

Πλέγμα	2x3	3x5	4x5	5x6	8x7
σφάλμα κβαντισμού	2,87E+01	1,11E+01	8,29E+00	5,51E+00	2,92E+00
εντροπία	2,02E-72	1,18E-64	2,92E-65	6,19E-67	1,28E-64
τοπογραφικό σφάλμα	3,33E-01	6,27E-01	4,67E-01	4,00E-01	6,33E-01
παραμόρφωση	9,19E+01	4,10E+01	3,03E+01	2,69E+01	1,26E+01
τοπογραφικό γινόμενο	-0,228	-0,434	-0,175	-0,457	-0,574

6.3.4.2 Πραγματικά σύνολα δεδομένων

Globins 1 (30 δείγματα)

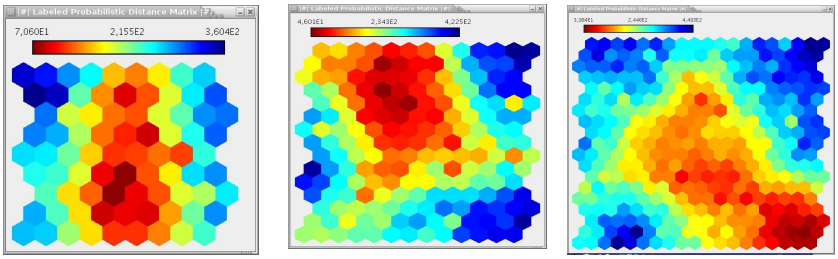
Πλέγμα	3x2	3x3	3x5	5x6	7x8	10x9
σφάλμα κβαντισμού	3,07E+01	2,51E+01	2,15E+01	1,46E+01	1,12E+01	3,07E+01
εντροπία	2,50E-142	2,30E-133	2,93E-135	1,94E-131	1,90E-132	2,50E-142
τοπογραφικό σφάλμα	3,33E-02	1,67E-01	1,33E-01	5,00E-02	1,00E-01	3,33E-02
παραμόρφωση	9,00E+01	7,53E+01	7,70E+01	5,82E+01	5,03E+01	9,00E+01
τοπογραφικό γινόμενο	-0,103	-0,433	-0,932	-0,264	-0,281	-0,266



Εικόνα 6.17 Πλέγματα 5x6, 7x8, 10x9 για το πρώτο πρωτεϊνικό σύνολο δεδομένων

Globins 2 (560 δείγματα)

Πλέγμα	3x2	3x5	5x5	7x8	10x10	15x15	28x20
σφάλμα κβαντισμού	2,49E+01	9,64E+00	5,71E+00	2,17E+00	1,38E+00	6,09E-01	2,42E-01
εντροπία	1,47E-13 2	2,24E-12 7	3,79E-12 7	2,25E-10 4	4,01E-12 3	3,04E-12 0	9,53E-12 2
τοπογραφικό σφάλμα	1,25E-02	3,04E-02	1,29E-01	1,00E-01	2,13E-01	6,18E-01	6,55E-01
παραμόρφωση	6,57E+01	3,51E+01	2,25E+01	8,69E+00	6,36E+00	2,91E+00	1,17E+00
τοπογραφικό γινόμενο	-0,0683	-0,248	-0,331	-0,301	-0,314	-0,339	ΔΥ



Εικόνα 6.18 Πλέγματα 5x5, 7x8, 10x10 για το δεύτερο πρωτεϊνικό σύνολο δεδομένων. Οι συστάδες απεικονίζονται με σκούρο μπλε

6.3.4.3 Συμπεράσματα

Παρατηρώντας τα αποτελέσματα των προσομοιώσεων μπορούμε να χωρίσουμε ανάλογα με την επίδοσή τους τα μέτρα οργάνωσης σε τρεις κατηγορίες:

- σφάλμα κβαντισμού και παραμόρφωση, μειώνονται μονότονα με την αύξηση του μεγέθους του πλέγματος.
- τοπογραφικό σφάλμα και τοπογραφικό γινόμενο. Αυξομειώνονται παρουσιάζοντας, συνήθως, κάποιο ελάχιστο ή μέγιστο αντίστοιχα για κάποιο πλέγμα, που δείχνει και το βέλτιστο, σύμφωνα με το κριτήριο πλέγμα.
- εντροπία, δεν παρουσιάζει κάποια επιθυμητή συμπεριφορά.

Για την πρώτη κατηγορία μέτρων παρατηρείται μείωση με την αύξηση του πλέγματος σε όλες τις περιπτώσεις συνόλων δεδομένων. Το γεγονός αυτό ήταν αναμενόμενο, αφού αφενός για το σφάλμα κβαντισμού περισσότεροι κόμβοι συνεπάγονται πιο συμπαγείς συστάδες και άρα μικρότερο συνολικό σφάλμα, ενώ για το μέτρο παραμόρφωσης με τον ίδιο τρόπο η αύξηση του μεγέθους του χάρτη οδηγεί σε μείωση της τιμής παραμόρφωσης Τα μέτρα αυτά θα μπορούσαν να χρησιμοποιηθούν από τον ερευνητή σε μια λογική tradeoff μεταξύ της επίδοσής τους και του κόστους (χρόνος και απαιτούμενοι πόροι) εκπαίδευσης του SOHMMM. Δεν είναι πάντως δυνατόν να μας δώσουν κάποιο ελάχιστο με βάση το οποίο να μπορούσαμε να προσδιορίσουμε το ιδανικό μέγεθος του χάρτη.

Η δεύτερη κατηγορία μέτρων είναι και η πιο σημαντική στην απόφαση σχετικά με το

βέλτιστο μέγεθος χάρτη. Σε ότι αφορά το τοπογραφικό σφάλμα, παρατηρούμε πως έχει τοπικό ελάχιστο για πλήθος κόμβων ίσο με το πλήθος των δεδομένων εισόδου σε όλες τις περιπτώσεις εκτός του δεύτερου πρωτεϊνικού dataset. Τονίζεται σε αυτό το σημείο πως πρέπει να αγνοηθούν οι υπερβολικά καλές τιμές που παρατηρούνται σε όλες τις περιπτώσεις για πολύ μικρό πλέγμα (3x2) και που οφείλονται σε κακή συμπεριφορά του μέτρου λόγω του μικρού μεγέθους (πχ για πλέγμα 3x2 μόνο οι κόμβοι στις άκρες δεν θα έχουν γειτονικούς όλους τους άλλους, συνεπώς οι παραβιάσεις για πρώτη και δεύτερη BMU γειτονικές θα είναι ελάχιστες).

Το τοπογραφικό γινόμενο παρουσιάζει μέγιστο για πλήθος κόμβων ίσο με το πλήθος των δεδομένων εισόδου στις περιπτώσεις των δύο πρώτων τεχνητών δεδομένων εισόδου και του πρώτου πρωτεϊνικού. Στην περίπτωση του δεύτερου πρωτεϊνικού dataset δεν υπολογίστηκε η τιμή του τοπογραφικού γινομένου για πλέγμα ίσο με τη διάσταση των δεδομένων, επειδή αυτό απαιτούσε υπολογισμούς η χρονική διάρκεια των οποίων ήταν πάρα πολύ μεγάλη. Πάντως σε αυτή την περίπτωση το τοπογραφικό γινόμενο δείχνει να προκρίνει ως καλύτερο πλέγμα το 7x8 συμφωνώντας με το τοπογραφικό σφάλμα. Σε αυτό το σημείο πρέπει να σημειωθεί ότι, τουλάχιστον για τα πειράματα που εκτελέστηκαν, το τοπογραφικό γινόμενο δεν αποκτά ποτέ τιμή 0 (τέλεια τοπολογία) ή θετική τιμή. Αντίθετα, είναι πάντοτε αρνητικό, προσεγγίζοντας το 0 για καλές απεικονίσεις. Επίσης, αξίζει να αναφερθεί ότι το τοπογραφικό γινόμενο έχει πολύ μεγαλύτερες υπολογιστικές απαιτήσεις σε σχέση με τα άλλα μέτρα που καθιστούν τη χρήση του μη επιθυμητή σε πολλές περιπτώσεις (πχ όπως για το δεύτερο πρωτεϊνικό dataset). Πάντως δείχνει να είναι σε θέση να περιγράψει με περισσότερη ακρίβεια την ποιότητα της οργάνωσης από τοπολογικής απόψεως.

Η εντροπία τέλος, έχει θετική συμπεριφορά στην περίπτωση του δευτέρου τεχνητού και του δευτέρου πρωτεϊνικού dataset, πάντως η γενική συμπεριφορά της δεν κρίνεται αξιόπιστη, οπότε δεν προκρίνεται η επιλογή της για τον προσδιορισμό του μεγέθους πλέγματος.

Εν κατακλείδι, θα μπορούσαμε να προτείνουμε την χρήση των τεσσάρων μέτρων (εκτός εντροπίας) για την απόφαση του βέλτιστου μεγέθους για το πλέγμα, σε συνάρτηση πάντως με τις αναφερθείσες παρατηρήσεις. Προφανώς, τα συμπεράσματα πρέπει πάντοτε να λαμβάνονται υπόψιν με επιφύλαξη και με τη βοήθεια κάποιας οπτικοποίησης.

6.4 Σύνοψη

Στο κεφάλαιο αυτό μελετήθηκε η επίδοση των μέτρων οργάνωσης σε τέσσερις διαφορετικές λειτουργίες: μονοτονία κατά την διάρκεια της εκπαίδευσης, επιλογή αριθμού εποχών, αριθμού καταστάσεων των HMM και μεγέθους πλέγματος. Η διερεύνηση αυτή προφανώς και δεν εξαντλήθηκε αφού θα μπορούσαμε να πραγματοποιήσουμε κάποια μελέτη σε σχέση και με τις υπόλοιπες παραμέτρους του SOHMMM ή να εμβαθύνουμε παραπάνω σε κάποια μέτρα ή επιμέρους λειτουργίες. Στόχος μας πάντως δεν ήταν η εξάντλητική παρουσίαση των δυνατοτήτων των μέτρων οργάνωσης που αναπτύξαμε, αλλά να καταδείξουμε ότι τα μέτρα αυτά είναι σε θέση να ποσοτικοποιήσουν ικανοποιητικά τη διαδικασία και τη διάταξη ενός SOHMMM καθώς και να χρησιμοποιηθούν σε ποίκιλες λειτουργίες. Εν τέλει, κάποια μέτρα αποδείχθηκαν καταλληλότερα να περιγράψουν συγκεκριμένες παραμέτρους της διαδικασίας οργάνωσης σε σχέση με άλλα, με όλα παντως να έχουν κάποια δυνατά/αδύνατα σημεία. Για αυτό τον λόγο, δεν θελήσαμε να προχωρήσουμε σε μια αυστηρή αξιολόγησή τους, αλλά με βάση την εκάστοτε μελέτη καταδείξαμε από ποια άποψη ένα μέτρο μπορούσε να είναι καλύτερο από ένα άλλο.

Τελικά, καταδείχθηκε ένα σημαντικό μέρος των δυνατοτήτων των μέτρων της

οργάνωσης και της ποιότητας του SOHMMM που αναπτύξαμε, δείχνοντας πώς αυτά είναι ικανά να ποσοτικοποιήσουν διαφορετικές λειτουργίες/παραμετροποιήσεις στην διαδικασία εκπαίδευσης του Self Organizing Hidden Markov Model Map.

Βιβλιογραφία

Κεφάλαιο 2

[1]T. Kohonen, M. R. Schroeder, and T. S. Huang, "*Self-Organizing Maps*", 3rd. Springer-Verlag New York, 2001

[2] http://www.cis.hut.fi/research/som-bibl/

[3] http://www.cis.hut.fi/projects/somtoolbox/

[4] Y. Cheng, "*Convergence and ordering of kohonens batch map"*, Neural Computation, vol. 9(8), p.1667,1997

[5] F. Mulier, V. Cherkassky, in Proc. 12th IAPR, Int. Conf. on Pattern Recognition, IEEE Service Center, Piscataway, NJ 1994, p. 11-224

[6] E. Erwin, K. Obermayer, K. Schulten: Biol. Cyb. 67, 35, 1992

[7] J. W. Sammon Jr., A nonlinear mapping for data structure analysis, IEEE Trans. Comp. C-18,p. 401, 1969

[8] H-U Bauer and K. Pawelzik, Quantifying neighborhood preservation in SOMs, IEEE Transactions, 1992

[9] S. Zrehen: In Proc. ICANN'93, Int. Conf. on Artificial Neural Networks, Springer-Verlag, London 1993, p. 609

[10] T. Villmann, R. Der, T. Martinetz: In Proc. ICNN'94, IEEE Int. Conf on Neural Networks, IEEE Service Center, Piscataway, NJ, 1994, p. 645

[11] K. Kiviluoto, Proc. ICNN'96, IEEE Int. Conf. on Neural Networks, IEEE Service Center, Piscataway, NJ 1996, p. 294

[12] S. Kaski, K. Lagus, Lecture Notes in Computer Science, vol. 1112, ed. by C. v. d. Malsburg, W. von Seelen, J. C. Vorbruggen, B. Sendhoff, Springer, Berlin 1996, p. 809

[13] A. Ultsch, H. Siemon: Technical Report 329, Univ. of Dortmund, Dort-Dortmund, Germany 1989

[14] M. A. Kraaijveld, J. Mao, A. K. Jain: In Proc. 11 IAPR, Int. Conf. On Pattern Recognition, IEEE Comput. Soc. Press, Los Alamitos, CA, 1992, p. 41

Κεφάλαιο 3

[1] Lawrence R. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of the IEEE,vol. 77,no. 2,Feb. 1989

[2] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state

Markov chains,”Ann. Math. Stat., vol 37,pp. 1554-1563,1966

[3] L. E. Baum and J. A. Egon, ”An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology,” Bull. Amer. Meteorol. Soc., vol. 73, pp. 360-363, 1967.

[4] L. E. Baum and G. R. Sell, “Growth functions for transformations on manifolds,” Pac. /. Math., vol. 27, no. 2, pp. 211-227, 1968.

[5] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, “A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains,” Ann. Math. Stat., vol. 41, no. 1, pp. 164-171, 1970.

[6] L. E. Baum, “An inequalityand associated maximization technique in statistical estimation for probabilistic functions of Markov processes,” Inequalities, vol. 3, pp. 1-8, 1972.

[7] J. K. Baker, ”The dragon system-An overview,” IEEE Trans. Acoust. Speech Signal Processing, vol. ASP-23, no. 1, pp. 24-29, Feb. 1975.

[8] F. Jelinek, “A fast sequential decoding algorithm using a stack,” ISM]. Res. Develop., vol. 13, pp. 675-685, 1969.

[9] L. R. Bahl and F. Jelinek, “Decoding for channels with insertions, deletions, and substitutions with applications to speech recognition,” IEEE Trans. Informat. Theory, vol IT-21, pp. 404-411, 1975.

[10] F. Jelinek, L. R. Bahl, and R. L. Mercer, ”Design of a linguistic statistical decoder for the recognition of continuous speech, ”IEEE Trans. Informat. Theory, vol. IT-21, pp. 250-256, 1975.

[11] F. Jelinek, “Continuous speech recognition by statistical methods,” Proc. IEEE , vol. 64, pp. 532-536, Apr. 1976.

[12] R. Bakis, ”Continuous speech word recognition via centisecond acoustic states,” in Proc. ASA Meeting, Washington, DC, Apr. 1976.

[13] F. Jelinek, L. R. Bahl, and R. L. Mercer, “Continuous speech recognition: Statistical methods,” in Handbook of Statistics, P. R. Krishnaiad, Ed. Amsterdam,The Netherlands: North- Holland, 1982.

[14] L. R. Bahl, F. Jelinek, and R. L. Mercer, “A maximum likelihood approach to continuous speech recognition,” IEEE Trans. Pattern Anal. Machine Intel., vol. PAMI-5, pp. 179-190, 1983.

[15] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, “An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition,” SeIISyst. Tech. /., vol. 62, no. 4, pp. 1035-1074, Apr. 1983.

[16] B. H. Juang,“On the hidden Markov model and dynamic time warping for speech recognition-A unified view,” AT&TTech..,vol. 63, no. 7, pp. 1213-1243, Sept. 1984.

[17] L. R. Rabiner and B. H. Juang,"An introduction to hidden Markov models," IEEE ASSP Mag., vol. 3, no. l, pp. 4-16, 1986.

[18] J. S. Bridle, "Stochastic models and template matching: Some important relationships between two apparently different techniques for automatic speech recognition," in Proc. Inst. ofAcoustics, Autum Conf., pp. 1-8, Nov. 1984.

[19] J. Makhoul, S. Roucos, and H. Gish, "Vector quantization in speech coding,"Proc. IEEE vol. 73, no. 11, pp. 1551-1588, Nov. 1985.

[20] S. E. Levinson, "Structural methods in automatic speech recognition," Proc. IEEE,vol. 73, no. 11, pp. 1625-1650, Nov. 1985.

[21] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimal decoding algorithm," IEEE Trans. Informat. Theory, vol. IT-13, pp. 260-269, Apr. 1967.

[22] G. D. Forney, "The Viterbi algorithm," Proc. IEE, vo l. 61, pp. 268-278, Mar. 1973.

[23] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," 1. Roy. Stat. Soc., vol. 39, no. 1, pp. 1-38, 1977.

[24] L. A. Liporace, "Maximum likelihood estimation for multivariate observations of Markov sources," IEEE Trans. Lnformat. Theory, vol. IT-28, no. 5, pp. 729-734, 1982.

[25] B. H. Juang, "Maximum likelihood estimation for mixture multivariate stochastic observations of Markov chains," AT&T Tech. I., vol. 64, no. 6, pp. 1235-1249, July-Aug. 1985.

[26] B. H. Juang, S. E. Levinson, and M. M. Sondhi, "Maximum likelihood estimation for multivariate mixture observations of Markov chains," /IEEE Trans. lnformat. Theory, vol. IT-32, no. 2, pp. 307-309, Mar. 1986.

[27] A. B. Poritz, "Linear predictive hidden Markov models and the speech signal," in Proc. lCASSP '82 (Paris, France), pp. 1291-1294, May 1982.

[28] B. H. Juang and L. R. Rabiner, "Mixture autoregressive hidden Markov models for speech signals," IEEE Trans. Acoust. Speech Signal Processing, vol. ASSP-33, no. 6, pp. 1404-1413, Dec. 1985.

[29] B. Lowerreand R. Reddy, "The HARPY speech understanding system," in Trends in Speech Recognition, W. Lea, Editor. Englewood Cliffs, NI: Prentice-Hall, 1980, pp. 340-346.

[30] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in Proc. LCASSP '86 (Tokyo, Japan), pp. 49-52, Apr. 1986.

[31] Y. Ephraim, A. Dembo, and L. R. Rabiner, "A minimum discrimination information approach for hidden Markov modeling," in Proc. lCASSP '87(Dallas, TX), Apr. 1987.

[32] B. H. Juang and L. R. Rabiner, “A probabilistic distance measure for hidden Markov models,” AT&T Tech. /., vol. 64, no. 2, pp. 391-408, Feb. 1985

Κεφάλαιο 4

[1] C. Ferles and A.-G. Stafylopatis, An Overview of the Self-Organizing Hidden Markov Model Map, to appear in IEEE

[2] P. Baldi and S. Brunak, *Bioinformatics: The Machine Learning Approach*, 2nd ed. Cambridge, Massachusetts: The MIT Press, 2001.

[3] J.-S. R. Sang, C.-T. Sun and E. Mizutami, *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*. Upper Saddle River, New Jersey: Prentice-Hall, 1997.

[4] P. Baldi, “Gradient descent learning algorithms overview: a general dynamical systems perspective,” *IEEE Trans. Neural Networks*, vol.6, pp. 182-195, Jan. 1995.

[5] P. Baldi and Y. Chauvin, “Smooth on-line learning algorithms for hidden Markov models,” *Neural Computation*, vol. 6, pp. 305-316, March 1994.

[6] T. Kohonen, *Self-Organizing Maps*, 3rd ed. Berlin: Springer-Verlag, 2001.

[7] L. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proc. IEEE*, vol. 77, pp. 257-286, Feb. 1989.

Κεφάλαιο 5

[1] Daniel Polani, Measures for the organization of self-organizing maps, Self-organizing Neural Networks: Recent Advances and Applications, Physica-Verlag, 2001

[2] T. Villmann, R. Der, M. Hermann, T. Martinetz, Topology Preservation in Self-Organizing Feature Maps: Exact Definition and Measurement. IEEE Trans. Neural Networks 8(2), p. 256-266, 1997

[3] Goodhill, Quantifying neighborhood preservation in topographic mappings, Proceedings of the 3rd Joint Symposium on Neural Computation, 1996

[4] H. Ritter, K. Shulten, Convergence properties of Kohonen's Topology Conserving Maps:Fluctuation, Stability and Dimension Selection, Biological Cybernetics 60, p. 59-71, 1989

[5] R. Der, M. Hermann, Phase Transitions in Self-Organized Maps,Proc ICANN'93, International Conference on Artificial Neural Networks, p. 597, London, UK, Springer, 1993

[6] T. Graepel , M. Burger, K. Obermayer, Deterministic annealing for topographic vector

quantization and self-organizing maps, Proceedings of WSOM'97, p. 345-350, Espoo,Finland, University of Technology,Neural Networks Research Centre, 1997

[7] V. Hulle, Topology-preserving map formation achieved with a purely local unsupervised competitive learning rule, Neural Networks, 10(3), p. 431, 1997

[8] M.M. V. Hulle, Non parametric Denstiy Estimation and Regression Achieved with Topographic Maps Maximizing the Information-Theoretic Entropy of Their Outputs, Biological Cybernetics 77, p. 49-61, 1997

[9] M. M. V. Hulle, Fathful Representations and Topographic Maps, Wiley, 2000

[10] M. Alder, R. Togneri, Y. Attikiouzel, Dimension of the speech space, Proceedings of the IEE-1, 138(3), p. 207-214, 1991

[11] B. H. Juang and L. R. Rabiner, “A probabilistic distance measure for hidden Markov models,” AT&T Tech. /., vol. 64, no. 2, pp. 391-408, Feb. 1985

[12] H.-U. Bauer, M. Hermann, T. Vilmann, Neural maps and topographic vector quantization, Neural Networks 12, vol 4-5, p. 659-676, 1999

[13] R. Linsker, Self-Organization in a Perceptual Network. Computer 21, vol. 3, p. 105-117, 1988

[14] S. Haykin, Neural Networks:a comprehensive foundation, Prentice Hall, 1999

[15] Kiviluoto, J. Topology preservation in selforganizing maps. In Proceedings of the International Conference on Neural Networks (ICNN’96), 1:294–299, Piscataway, New Jersey, USA, IEEE Neural Networks Council, June 1996.

[16] G. H. Hardy, J. E. Littlewood, G. Polya, Inequalities. p. 261 Cambridge University Press, 1934

[17] R. E. Burkard, Quadratic assignment problems, Europ. Journ. Oper. Res., vol 15, p. 283-289, 1984

[18] E. L. Lawler, The quadratic assignment problem. Management Science, vol 9, p. 586-599. 1963

[19] G. Finke, R.E. Burkard, F. Rendl, Quadratic assignment problems. Annals of Discrete Mathematics , vol 31 , p. 61-82, 1987

[20] P Demartines, J Herault, Curvilinear Component Analysis, Quinzieme Coloque Gretsi, Juan-les-Pins, France, 1995

[21] J. W. Sammon Jr., A nonlinear mapping for data structure analysis, IEEE Trans. Comp. C-18,p. 401, 1969

[22] J Lampinen, Clustering Properties of Hierarchical Self-Organizing Maps, Journal of Mathematical Imaging and Vision 2, 261-272,1992

[23] Juha Vesanto, Mika Sulkava, Jaakko Hollmen, On the Decomposition of the Self-Organizing Map Distortion Measure, 2003

[24] Teuvo Kohonen. Self-Organizing Maps: Optimization Approaches, volume 2, p. 981–990, Elsevier Science Publishers, 1991

[25] Thore Graepel, Matthias Burger, and Klaus Obermayer. Phase transitions in stochastic self-organizing maps. Physical Review E , 56:3876–3890, 1997

[26] H-U Bauer and K. Pawelzik, Quantifying neighborhood preservation in SOMs, IEEE Transactions, 1992

[27] W. Liebert, K. Pawelzik, H. G. Schuster, Optimal Embeddings of Chaotic Attractors from Topological Considerations, Europhys. Lett. 14, p. 521-526, 1991

[28] T. Caliński, J. Harabasz, A dendrite method for cluster analysis', Communications in Statistics - Theory and Methods, 3:1, p. 1 – 27, 1974

[29] G. Milligan and M. Cooper, “An examination of procedures for determining the number of clusters in a data set,” Psychometrika, vol. 50, pp. 159–179, 1985

[30] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Trans. Pattern Recognition Machine Intell. 1 (2), p. 224–227, 1979

[31] J. Dunn, Well separated clusters and optimal fuzzy partitions, J. Cybernet. 4, p.95–104, 1974

[32] L. Hubert, J. Schultz, Quadratic assignment as a general data-analysis strategy, Br. J. Math. Statist. Psychol. 29, p. 190–241,1976

[33] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20, p. 53–65, 1987

[34] Ball, G.H., & Hall, D.J. . ISODATA, A novel method of data analysis and pattern classification (Tech. Rep. NTIS No. AD 699616). Menlo Park, CA: Stanford Research Institute, 1965

[35] D. Gusfield, Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge, U.K.: Cambridge Univ. Press, 1997.

[36] D. Sankoff and J. Kruskal, Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison. Stanford, CA: CSLI Publications, 1999.

[37] Jarkko Venna, Samuel Kaski, Neighborhood Preservation in Nonlinear Projection Methods: An Experimental Study

[38] S. Kaski, K. Lagus, Lecture Notes in Computer Science, vol. 1112, ed. by C. v. d. Malsburg, W. von Seelen, J. C. Vorbruggen, B. Sendhoff (Springer, Berlin 1996), p. 809

[39] Rui Xu, Donald Wunsch II, Survey of Clustering Algorithms, IEEE Transactions on Neural Networks, vol. 16, no. 3, May 2005

Κεφάλαιο 6

[1] A.Krogh, M. Brown, I. S. Mian, K. Sjolander and D. Haussler, "Hidden Markov models in computational biology: applications to protein modeling," *J. Molecular Biology*, vol. 235, pp. 1501-1531, Feb. 1994.
[2] The iProClass Protein Knowledgebase (release 3.48). [Online]. Available: http://pir.georgetown.edu/

Παράρτημα

Παράρτημα Α

Από την κανονικοποιημένη εκθετική παραμετροποίηση της (4.) έπεται ότι οι μερικές παράγωγοι των παραμέτρων μετάβασης a_{ij} είναι

$$\frac{\partial a_{ij}}{\partial w_{ij}} = \frac{\frac{\partial e^{w_{ij}}}{\partial w_{ij}} \sum_{l=1}^{N} e^{w_{il}} - e^{w_{ij}} \frac{\partial \sum_{l=1}^{N} e^{w_{il}}}{\partial w_{ij}}}{\left(\sum_{l=1}^{N} e^{w_{il}}\right)^2} = \frac{e^{w_{ij}} \sum_{l=1}^{N} e^{w_{il}} - e^{w_{ij}} \lambda e^{w_{ij}}}{\left(\sum_{l=1}^{N} e^{w_{il}}\right)^2}$$

$$= \frac{e^{w_{ij}}}{\sum_{l=1}^{N} e^{w_{il}}} \left(1 - \frac{e^{w_{ij}}}{\sum_{l=1}^{N} e^{w_{il}}}\right) = a_{ij}(1-a_{ij}) \quad \text{(A.1)}$$

$$\frac{\partial a_{ij}}{\partial w_{ik}} = \frac{-e^{w_{ij}} \frac{\partial \sum_{l=1}^{N} e^{w_{il}}}{\partial w_{ik}}}{\left(\sum_{l=1}^{N} e^{w_{il}}\right)^2} = \frac{-e^{w_{ij}} \lambda e^{w_{ik}}}{\left(\sum_{l=1}^{N} e^{w_{il}}\right)^2}$$

$$= -\lambda \frac{e^{w_{ij}}}{\sum_{l=1}^{N} e^{w_{il}}} \frac{e^{w_{ik}}}{\sum_{l=1}^{N} e^{w_{il}}} = -a_{ij}a_{ik}, \qquad k \neq j \quad \text{(A.2)}$$

ομοίως για τις παραμέτρους εκπομπής $b_j(t)$ έχουμε

$$\frac{\partial b_j(t)}{\partial r_{jt}} = b_j(t)(1-b_j(t)) \qquad \text{(A.3)}$$

$$\frac{\partial b_j(t)}{\partial r_{jk}} = -b_j(t)b_j(k), \qquad k \neq t \quad \text{(A.4)}$$

και για τις αρχικές παραμέτρους κατάστασης π_j

$$\frac{\partial \pi_j}{\partial u_j} = \pi_j(1-\pi_j) \qquad \text{(A.5)}$$

$$\frac{\partial \pi_j}{\partial u_k} = -\pi_j \pi_k, \qquad k \neq j \qquad \text{(A.6)}$$

Για ευκολότερη αναφορά, υπενθυμίζουμε την αναδρομή για τις πρόσθιες και οπίσθιες πιθανότητες

$$\alpha_1(j) = \pi_j b_j(o_1), \qquad 1 \leq j \leq N \qquad \text{(A.7)}$$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(o_{t+1}), \quad 1 \leq t \leq T\text{-}1$$

$$1 \leq j \leq N \qquad \text{(A.8)}$$

$$\beta_T(j) = 1, \qquad 1 \leq i \leq N \qquad \text{(A.9)}$$

$$\beta_t(j) = \sum_{i=1}^{N} a_{ji} b_i(o_{t+1}) \beta_{t+1}(i), \qquad t = T\text{-}1, T\text{-}2, \ldots, 1$$

$$1 \leq j \leq N \qquad \text{(A.10)}$$

Λόγω της ανάλυσης που έχει γίνει για το πρόβλημα της αποτίμησης χρησιμοποιώντας τις πρόσθιες και οπίσθιες μεταβλητές, υπάρχουν T δυνατοί τρόποι υπολογισμού της $P(O \mid \lambda)$

$$P(O \mid \lambda) = \sum_{j=1}^{N} \alpha_t(j) \beta_t(j), \qquad t = 1, 2, \ldots, T \qquad \text{(A.11)}$$

Για $t = T$ χρησιμοποιούμε τον τύπο αποτίμησης στη μορφή

$$P(O \mid \lambda) = \sum_{j=1}^{N} \alpha_T(j) \beta_T(j)$$

στη συνέχεια χρησιμοποιούμε την εμπρόσθια αναδρομή (A.8) για τις $\alpha_T(j)$ λαμβάνοντας

$$P(O \mid \lambda) = \sum_{j=1}^{N} \sum_{i=1}^{N} \alpha_{T-1}(i) a_{ij} b_j(o_T) \beta_T(j) \qquad \text{(A.12)}$$

Λήμμα 1:

$$\frac{\partial P(O \mid \lambda)}{\partial a_{rs}} = \sum_{l=1}^{T-1} \alpha_l(r) b_s(o_{l+1}) \beta_{l+1}(s) \qquad \text{(A.13)}$$

Απόδειξη: Παραγωγίζουμε μερικώς την (A.12) σε σχέση με το a_{rs} και λαμβάνουμε με τον κανόνα παραγώγισης για γινόμενο

$$\frac{\partial P(O \mid \lambda)}{\partial a_{rs}} = \alpha_{T-1}(r) b_s(o_T) \beta_T(s) + \sum_{j=1}^{N} \sum_{i=1}^{N} \frac{\partial \alpha_{T-1}(i)}{\partial a_{rs}} a_{ij} b_j(O_T) \beta_T(j) . \qquad \text{(A.14)}$$

Συνεπώς, χρειάζεται μόνο να υπολογίσουμε τον δεύτερο όρο στην δεξιά πλευρά της εξισωσης, κάτι που επιτυγχάνεται χρησιμοποιώντας μια ακόμα φορά την εξίσωση (A.8)

$$\frac{\partial \alpha_{T-1}(i)}{\partial a_{rs}} = \frac{\partial}{\partial a_{rs}} \left[\sum_{l=1}^{N} \alpha_{T-2}(l) a_{li} \right] b_i(o_{T-1}) \qquad \text{(A.15)}$$

διαφορίζοντας λαμβάνουμε

$$\frac{\partial \alpha_{T-1}(s)}{\partial a_{rs}} = \alpha_{T-2}(r) b_s(o_{T-1}) + \left[\sum_{l=1}^{N} \frac{\partial \alpha_{T-2}(l)}{\partial a_{rs}} a_{ls} \right] b_s(o_{T-1})$$

$$\frac{\partial \alpha_{T-1}(i)}{\partial a_{rs}} = \left[\sum_{l=1}^{N} \frac{\partial \alpha_{T-2}(l)}{\partial a_{rs}} a_{li} \right] b_i(o_{T-1}) , \qquad i \neq s$$

Εισαγωγή αυτών των μερικών παραγώγων στον δεύτερο όρο της (A.14) δίνει

$$\begin{aligned} \sum_{j=1}^{N} \sum_{i=1}^{N} \frac{\partial \alpha_{T-1}(i)}{\partial a_{rs}} a_{ij} b_j(O_T) \beta_T(j) \\ &= \sum_{j=1}^{N} \alpha_{T-2}(r) b_s(o_{T-1}) a_{sj} b_j(O_T) \beta_T(j) \\ &+ \sum_{j=1}^{N} \left[\sum_{l=1}^{N} \frac{\partial \alpha_{T-2}(l)}{\partial a_{rs}} a_{ls} \right] b_s(o_{T-1}) a_{sj} b_j(O_T) \beta_T(j) \\ &+ \sum_{j=1}^{N} \sum_{i=1, i \neq s}^{N} \left[\sum_{l=1}^{N} \frac{\partial \alpha_{T-2}(l)}{\partial a_{rs}} a_{li} \right] b_i(o_{T-1}) a_{ij} b_j(O_T) \beta_T(j) . \qquad \text{(A.16)} \end{aligned}$$

Και τοποθετώντας τους δύο τελευταίους όρους της (A.16) μαζί, λαμβάνουμε

$$\begin{aligned} \sum_{j=1}^{N} \sum_{i=1}^{N} \frac{\partial \alpha_{T-1}(i)}{\partial a_{rs}} a_{ij} b_j(O_T) \beta_T(j) \\ &= \sum_{j=1}^{N} \alpha_{T-2}(r) b_s(o_{T-1}) a_{sj} b_j(O_T) \beta_T(j) \\ &+ \sum_{j=1}^{N} \sum_{i=1}^{N} \sum_{l=1}^{N} \frac{\partial \alpha_{T-2}(l)}{\partial a_{rs}} a_{li} b_i(o_{T-1}) a_{ij} b_j(O_T) \beta_T(j) . \qquad \text{(A.17)} \end{aligned}$$

Στη συνέχεια, υπολογίζουμε τις τιμές κάθε ενός από τους δύο όρους στην δεξιά πλευρά της (A.17). Αρχικά,

$$\sum_{j=1}^{N} \alpha_{T-2}(r) b_s(o_{T-1}) a_{sj} b_j(O_T) \beta_T(j) = \alpha_{T-2}(r) b_s(o_{T-1}) \sum_{j=1}^{N} a_{sj} b_j(O_T) \beta_T(j)$$

με χρήση της οπίσθιας αναδρομής (A.10), λαμβάνουμε

$$\alpha_{T-2}(r) b_s(o_{T-1}) \sum_{j=1}^{N} a_{sj} b_j(O_T) \beta_T(j) = \alpha_{T-2}(r) b_s(o_{T-1}) \beta_{T-1}(s). \quad \text{(A.18)}$$

Δεύτερον, εναλλάσσοντας την σειρά των αθροίσεων στον δεύτερο όρο της (A.17)

$$\sum_{j=1}^{N} \sum_{i=1}^{N} \sum_{l=1}^{N} \frac{\partial \alpha_{T-2}(l)}{\partial a_{rs}} a_{li} b_i(o_{T-1}) a_{ij} b_j(O_T) \beta_T(j)$$
$$= \sum_{i=1}^{N} \sum_{l=1}^{N} \frac{\partial \alpha_{T-2}(l)}{\partial a_{rs}} a_{li} b_i(o_{T-1}) \sum_{j=1}^{N} a_{ij} b_j(O_T) \beta_T(j)$$

όμως, από την οπίσθια αναδρομή της (A.10) λαμβάνεται

$$\sum_{i=1}^{N} \sum_{l=1}^{N} \frac{\partial \alpha_{T-2}(l)}{\partial a_{rs}} a_{li} b_i(o_{T-1}) \sum_{j=1}^{N} a_{ij} b_j(O_T) \beta_T(j)$$
$$= \sum_{i=1}^{N} \sum_{l=1}^{N} \frac{\partial \alpha_{T-2}(l)}{\partial a_{rs}} a_{li} b_i(o_{T-1}) \beta_{T-1}(i) . \quad \text{(A.19)}$$

Με χρήση των (A.17) - (A.19) ξαναγράφουμε την δεξιά πλευρά της (A.14) ως

$$\frac{\partial P(O|\lambda)}{\partial a_{rs}} = \alpha_{T-1}(r) b_s(o_T) \beta_T(s)$$
$$+ \alpha_{T-2}(r) b_s(o_{T-1}) \beta_{T-1}(s)$$
$$+ \sum_{i=1}^{N} \sum_{l=1}^{N} \frac{\partial \alpha_{T-2}(l)}{\partial a_{rs}} a_{li} b_i(o_{T-1}) \beta_{T-1}(i) . \quad \text{(A.20)}$$

Όμως, το διπλό άθροισμα έχει ακριβώς την ίδια μορφή με αυτή που υπολογίστηκε στην εξίσωση (A.14). Συνεπώς, μπορούμε να συνεχίσουμε τη διαφόριση της $\frac{\partial \alpha_{T-2}(l)}{\partial a_{rs}}$ και να οργανώσουμε τις προκύπτουσες εκφράσεις όμοια με παραπάνω. Εξαιτίας του γεγονότος ότι το $\alpha_1(j)$ δεν περιέχει κάποιο a_{rs}, γίνεται φανερό ότι η διαφόριση τερματίζει στο σημείο όπου εμφανίζεται $\frac{\partial \alpha_1(l)}{\partial a_{rs}}$, αφού $\frac{\partial \alpha_1(l)}{\partial a_{rs}} = 0$. Επομένως, η (A.13) αποδεικνύεται.

Λήμμα 2:

$$\frac{\partial P(O \mid \lambda)}{\partial w_{ij}} = a_{ij} \sum_{l=1}^{T-1} \left[\alpha_l(i) b_j(o_{l+1}) \beta_{l+1}(j) - \alpha_l(i) \beta_l(i) \right] \qquad \text{(A.21)}$$

Απόδειξη: Χρησιμοποιώντας τον κανόνα της αλυσίδας για την αριστερή πλευρά της (A.21), λαμβάνουμε

$$\frac{\partial P(O \mid \lambda)}{\partial w_{ij}} = \sum_{z=1}^{N} \frac{\partial P(O \mid \lambda)}{\partial a_{iz}} \frac{\partial a_{iz}}{\partial w_{ij}} = \frac{\partial P(O|\lambda)}{\partial a_{ij}} \frac{\partial a_{ij}}{\partial w_{ij}} + \sum_{z=1, z \neq j}^{N} \frac{\partial P(O|\lambda)}{\partial a_{iz}} \frac{\partial a_{iz}}{\partial w_{ij}}$$

με εφαρμογή των (A.1), (A.2) και (A.13) στην δεξιά πλευρά της παραπάνω έκφρασης, καταλήγουμε στο

$$\begin{aligned}
\frac{\partial P(O|\lambda)}{\partial w_{ij}} &= \\
a_{ij}(1-a_{ij}) \sum_{l=1}^{T-1} \alpha_l(i) b_j(o_{l+1}) \beta_{l+1}(j) &- \sum_{z=1, z \neq j}^{N} a_{iz} a_{ij} \sum_{l=1}^{T-1} \alpha_l(i) b_z(o_{l+1}) \beta_{l+1}(z) \\
&= a_{ij} \sum_{l=1}^{T-1} \alpha_l(i) b_j(o_{l+1}) \beta_{l+1}(j) \\
&- a_{ij} \left[a_{ij} \sum_{l=1}^{T-1} \alpha_l(i) b_j(o_{l+1}) \beta_{l+1}(j) \right] \\
&- a_{ij} \sum_{z=1, z \neq j}^{N} a_{iz} \sum_{l=1}^{T-1} \alpha_l(i) b_z(o_{l+1}) \beta_{l+1}(z) \qquad \text{(A.22)}
\end{aligned}$$

Τοποθετώντας τους δύο τελευταίους όρους της δεξιάς πλευράς της (A.22) μαζί, λαμβάνουμε

$$\frac{\partial P(O|\lambda)}{\partial w_{ij}} = a_{ij} \left[\sum_{l=1}^{T-1} \alpha_l(i) b_j(o_{l+1}) \beta_{l+1}(j) - \sum_{z=1}^{N} a_{iz} \sum_{l=1}^{T-1} \alpha_l(i) b_z(o_{l+1}) \beta_{l+1}(z) \right] . \text{(A.23)}$$

Εναλλάσσοντας την σειρά των αθροίσεων στον δεύτερο όρο της (A.23) λαμβάνουμε

$$\sum_{z=1}^{N} a_{iz} \sum_{l=1}^{T-1} \alpha_l(i) b_z(o_{l+1}) \beta_{l+1}(z) = \sum_{l=1}^{T-1} \alpha_l(i) \sum_{z=1}^{N} a_{iz} b_z(o_{l+1}) \beta_{l+1}(z)$$

Αν χρησιμοποιήσουμε την οπίσθια αναδρομή (A.10), λαμβάνουμε

$$\sum_{l=1}^{T-1} \alpha_l(i) \sum_{z=1}^{N} a_{iz} b_z(o_{l+1}) \beta_{l+1}(z) = \sum_{l=1}^{T-1} \alpha_l(i) \beta_l(i) \qquad \text{(A.24)}$$

εισαγωγή της (A.24) στην (A.23) οδηγεί στην παρακάτω έκφραση

$$\frac{\partial P(O\mid\lambda)}{\partial w_{ij}} = a_{ij}\left[\sum_{l=1}^{T-1}\alpha_l(i)b_j(o_{l+1})\beta_{l+1}(j)-\sum_{l=1}^{T-1}\alpha_l(i)\beta_l(i)\right]$$
$$= a_{ij}\sum_{l=1}^{T-1}\left[\alpha_l(i)b_j(o_{l+1})\beta_{l+1}(j)-\alpha_l(i)\beta_l(i)\right]$$

οπότε αποδεικνύεται η (A.21).

Αρχικά, ορίζουμε την συνάρτηση δείκτη

$$I\{o_t=m|\lambda\}=\begin{matrix}1, & o_t=v_m\\ 0, & otherwise\end{matrix}$$

Λήμμα 3:

$$\frac{\partial P(O\mid\lambda)}{\partial b_x(y)} = \frac{1}{b_x(y)}\sum_{l=1}^{T}I\{o_l=y\mid\lambda\}\alpha_l(x)\beta_l(x) \qquad \text{(A.25)}$$

Απόδειξη: Διαφορίζουμε μερικώς την (A.12) ως προς την b_x(y), λαμβάνοντας από τον κανονα παραγώγισης για γινόμενο

$$\frac{\partial P(O\mid\lambda)}{\partial b_x(y)} = I\{o_T=y\mid\lambda\}\sum_{i=1}^{N}\alpha_{T-1}(i)a_{ix}\beta_T(x)$$
$$+\sum_{j=1}^{N}\sum_{i=1}^{N}\frac{\partial\alpha_{T-1}(i)}{\partial b_x(y)}a_{ij}b_j(o_T)\beta_T(j)$$

επιπλέον, χρήση της πρόσθιας αναδρομής (A.8) δίνει

$$\frac{\partial P(O\mid\lambda)}{\partial b_x(y)} = \frac{1}{b_x(o_T)}I\{o_T=y\mid\lambda\}\alpha_T(x)\beta_T(x)$$
$$+\sum_{j=1}^{N}\sum_{i=1}^{N}\frac{\partial\alpha_{T-1}(i)}{\partial b_x(y)}a_{ij}b_j(o_T)\beta_T(j). \qquad \text{(A.26)}$$

Συνεπώς, χρειάζεται να υπολογίσουμε μόνο τον δεύτερο όρο στην δεξιά πλευρά. Αυτό επιτυγχάνεται ξανά με την χρήση της εξίσωσης (A.8)

$$\frac{\partial\alpha_{T-1}(i)}{\partial b_x(y)} = \frac{\partial}{\partial b_x(y)}\left[\sum_{l=1}^{N}\alpha_{T-2}(l)a_{li}\right]b_i(o_{T-1}) \qquad \text{(A.27)}$$

Με διαφόριση:

$$\frac{\partial \alpha_{T-1}(x)}{\partial b_x(y)} = I\{o_{T-1} = y \mid \lambda\} \sum_{l=1}^{N} \alpha_{T-2}(l) a_{lx} + \left[\sum_{l=1}^{N} \frac{\partial \alpha_{T-2}(l)}{\partial b_x(y)} a_{lx} \right] b_x(o_{T-1})$$

$$\frac{\partial \alpha_{T-1}(i)}{\partial b_x(y)} = \left[\sum_{l=1}^{N} \frac{\partial \alpha_{T-2}(l)}{\partial b_x(y)} a_{li} \right] b_i(o_{T-1}), \qquad i \neq x$$

και εισαγωγή αυτών μερικών παραγώγων στον δεύτερο όρο της (A.26) δίνει

$$\begin{aligned}
&\sum_{j=1}^{N} \sum_{i=1}^{N} \frac{\partial \alpha_{T-1}(i)}{\partial b_x(y)} a_{ij} b_j(o_T) \beta_T(j) \\
&\quad = \sum_{j=1}^{N} I\{o_{T-1} = y \mid \lambda\} \sum_{l=1}^{N} \alpha_{T-2}(l) a_{lx} a_{xj} b_j(o_T) \beta_T(j) \\
&\quad + \sum_{j=1}^{N} \left[\sum_{l=1}^{N} \frac{\partial \alpha_{T-2}(l)}{\partial b_x(y)} a_{lx} \right] b_x(o_{T-1}) a_{xj} b_j(o_T) \beta_T(j) \\
&\quad + \sum_{j=1}^{N} \sum_{i=1, i \neq x}^{N} \left[\sum_{l=1}^{N} \frac{\partial \alpha_{T-2}(l)}{\partial b_x(y)} a_{li} \right] b_i(o_{T-1}) a_{ij} b_j(o_T) \beta_T(j). \qquad \text{(A.28)}
\end{aligned}$$

Και τοποθετώντας μαζί τους δύο τελευταίους όρους της (A.28) λαμβάνουμε

$$\begin{aligned}
&\sum_{j=1}^{N} \sum_{i=1}^{N} \frac{\partial \alpha_{T-1}(i)}{\partial b_x(y)} a_{ij} b_j(o_T) \beta_T(j) \\
&\quad = \sum_{j=1}^{N} I\{o_{T-1} = y \mid \lambda\} \sum_{l=1}^{N} \alpha_{T-2}(l) a_{lx} a_{xj} b_j(o_T) \beta_T(j) \\
&\quad + \sum_{j=1}^{N} \sum_{i=1}^{N} \sum_{l=1}^{N} \frac{\partial \alpha_{T-2}(l)}{\partial b_x(y)} a_{li} b_i(o_{T-1}) a_{ij} b_j(o_T) \beta_T(j). \qquad \text{(A.29)}
\end{aligned}$$

Στη συνέχεια, υπολογίζουμε τις τιμές κάθε ενός από τους δύο όρους της δεξιάς πλευράς της (A.29). Αρχικά,

$$\begin{aligned}
&\sum_{j=1}^{N} I\{o_{T-1} = y \mid \lambda\} \sum_{l=1}^{N} \alpha_{T-2}(l) a_{lx} a_{xj} b_j(o_T) \beta_T(j) \\
&\quad = \sum_{l=1}^{N} I\{o_{T-1} = y \mid \lambda\} \alpha_{T-2}(l) a_{lx} \sum_{j=1}^{N} a_{xj} b_j(o_T) \beta_T(j)
\end{aligned}$$

Χρησιμοποιώντας την οπίσθια αναδρομή (A.10), λαμβάνουμε

$$\sum_{l=1}^{N} I\{o_{T-1} = y \mid \lambda\} \alpha_{T-2}(l) a_{lx} \sum_{j=1}^{N} a_{xj} b_j(o_T) \beta_T(j) = \sum_{l=1}^{N} I\{o_{T-1} = y \mid \lambda\} \alpha_{T-2}(l) a_{lx} \beta_{T-1}(x)$$

και αν χρησιμοποιήσουμε την πρόσθια αναδρομή (A.8), λαμβάνουμε

$$\sum_{l=1}^{N} I\{o_{T-1} = y \mid \lambda\}\alpha_{T-2}(l)a_{lx}\beta_{T-1}(x)$$
$$= \frac{1}{b_x(o_{T-1})} I\{o_{T-1} = y \mid \lambda\}\alpha_{T-1}(x)\beta_{T-1}(x). \quad \text{(A.30)}$$

Δεύτερον, εναλλάσσοντας την σειρά των αθροίσεων στον δεύτερο όρο της (A.29) λαμβάνουμε

$$\sum_{j=1}^{N}\sum_{i=1}^{N}\sum_{l=1}^{N}\frac{\partial\alpha_{T-2}(l)}{\partial b_x(y)} a_{li}b_i(o_{T-1})a_{ij}b_j(o_T)\beta_T(j)$$
$$= \sum_{i=1}^{N}\sum_{l=1}^{N}\frac{\partial\alpha_{T-2}(l)}{\partial b_x(y)} a_{li}b_i(o_{T-1})\sum_{j=1}^{N}a_{ij}b_j(o_T)\beta_T(j)$$

όμως από την οπίσθια αναδρομή (A.10) λαμβάνουμε

$$\sum_{i=1}^{N}\sum_{l=1}^{N}\frac{\partial\alpha_{T-2}(l)}{\partial b_x(y)} a_{li}b_i(o_{T-1})\sum_{j=1}^{N}a_{ij}b_j(o_T)\beta_T(j)$$
$$= \sum_{i=1}^{N}\sum_{l=1}^{N}\frac{\partial\alpha_{T-2}(l)}{\partial b_x(y)} a_{li}b_i(o_{T-1})\beta_{T-1}(i). \quad \text{(A.31)}$$

Χρήση των (A.29) - (A.31) διαμορφώνει την (A.26) ως

$$\frac{\partial P(O \mid \lambda)}{\partial b_x(y)} = \frac{1}{b_x(o_T)} I\{o_T = y \mid \lambda\}\alpha_T(x)\beta_T(x)$$
$$+ \frac{1}{b_x(o_{T-1})} I\{o_{T-1} = y \mid \lambda\}\alpha_{T-1}(x)\beta_{T-1}(x)$$
$$+ \sum_{i=1}^{N}\sum_{l=1}^{N}\frac{\partial\alpha_{T-2}(l)}{\partial b_x(y)} a_{li}b_i(o_{T-1})\beta_{T-1}(i). \quad \text{(A.32)}$$

Όμως, το διπλό άθροισμα έχει ακριβώς την ίδια μορφή με αυτή που υπολογίστηκε στην εξίσωση (A.26). Συνεπώς, μπορούμε να συνεχίσουμε τη διαφόριση της $\frac{\partial\alpha_{T-2}(l)}{\partial b_x(y)}$ και να οργανώσουμε τις προκύπτουσες εκφράσεις όμοια με παραπάνω. Εξαιτίας του γεγονότος ότι το $\alpha_1(j) = \pi_j b_j(o_1)$ λαμβάνουμε

$$\sum_{i=1}^{N}\sum_{l=1}^{N}\frac{\partial\alpha_1(l)}{\partial b_x(y)} a_{li}b_i(o_2)\beta_2(i)$$
$$= \sum_{i=1}^{N}\sum_{l=1}^{N}\pi_l\frac{\partial b_l(o_1)}{\partial b_x(y)} a_{li}b_i(o_2)\beta_2(i)$$

$$= \sum_{i=1}^{N} \pi_x \frac{\partial b_x(o_1)}{\partial b_x(y)} a_{xi} b_i(o_2) \beta_2(i)$$

$$= I\{o_1 = y \mid \lambda\} \pi_x \sum_{i=1}^{N} a_{xi} b_i(o_2) \beta_2(i)$$

όμως, από την (A.10) λαμβάνουμε

$$I\{o_1 = y \mid \lambda\} \pi_x \sum_{i=1}^{N} a_{xi} b_i(o_2) \beta_2(i) = I\{o_1 = y \mid \lambda\} \pi_x \beta_1(x)$$

Από την (A.7) προκύπτει ότι

$$I\{o_1 = y \mid \lambda\} \pi_x \beta_1(x) = \frac{1}{b_x(o_1)} I\{o_1 = y \mid \lambda\} \alpha_1(x) \beta_1(x). \qquad \text{(A.33)}$$

Συνεπώς, από τις (A.32) και (A.33) γίνεται φανερό ότι η διαφόριση τερματίζει όταν εμφανίζεται η $\frac{\partial \alpha_1(l)}{\partial b_x(y)}$, αφού $\frac{\partial \alpha_1(l)}{\partial b_x(y)}$ είναι η τελευταία έκφραση που μπορεί να περιέχει τον όρο $b_x(y)$.Κατά συνέπεια,

$$\frac{\partial P(O \mid \lambda)}{\partial b_x(y)} = \sum_{l=1}^{T} \frac{1}{b_x(o_l)} I\{o_l = y \mid \lambda\} \alpha_l(x) \beta_l(x). \qquad \text{(A.34)}$$

Λόγω της έκφρασης $I\{o_l = y \mid \lambda\}$ οι μόνοι μη μηδενικοί όροι στην παραπάνω άθροιση είναι αυτοί για τους οποίους ο κανόνας $o_l = v_y$ αληθεύει. Οπότε, μια ισοδύναμη έκφραση με την (A.34) είναι

$$\frac{\partial P(O \mid \lambda)}{\partial b_x(y)}$$

$$= \sum_{l=1}^{T} \frac{1}{b_x(y)} I\{o_l = y \mid \lambda\} \alpha_l(x) \beta_l(x)$$

$$= \frac{1}{b_x(y)} \sum_{l=1}^{T} I\{o_l = y \mid \lambda\} \alpha_l(x) \beta_l(x)$$

Οπότε και αποδεικνύεται η (A.25).

Λήμμα 4:

$$\frac{\partial P(O \mid \lambda)}{\partial r_{jt}} = \sum_{l=1}^{T} \left[I\{o_l = t \mid \lambda\} \alpha_l(j) \beta_l(j) - b_j(t) \alpha_l(j) \beta_l(j) \right] \qquad \text{(A.35)}$$

Απόδειξη: Με χρήση του κανόνα της αλυσίδας στην αριστερή πλευρά της (A.21), λαμβάνουμε

$$\frac{\partial P(O\mid\lambda)}{\partial r_{jt}} = \sum_{z=1}^{M}\frac{\partial P(O\mid\lambda)}{\partial b_j(z)}\frac{\partial b_j(z)}{\partial r_{jt}} = \frac{\partial P(O\mid\lambda)}{\partial b_j(t)}\frac{\partial b_j(t)}{\partial r_{jt}} + \sum_{z=1,z\neq t}^{M}\frac{\partial P(O\mid\lambda)}{\partial b_j(z)}\frac{\partial b_j(z)}{\partial r_{jt}}$$

εφαρμόζοντας τις (A.3), (A.4) και (A.25) στην δεξιά πλευρά της ανωτέρω έκφρασης λαμβάνουμε

$$\begin{aligned}\frac{\partial P(O\mid\lambda)}{\partial r_{jt}} &= b_j(t)(1-b_j(t))\frac{1}{b_j(t)}\sum_{l=1}^{T} I\{o_l = t\mid\lambda\}\alpha_l(j)\beta_l(j)\\ &\quad - \sum_{z=1,z\neq t}^{M} b_j(z)b_j(t)\frac{1}{b_j(z)}\sum_{l=1}^{T} I\{o_l = z\mid\lambda\}\alpha_l(j)\beta_l(j)\\ &= \sum_{l=1}^{T} I\{o_l = t\mid\lambda\}\alpha_l(j)\beta_l(j)\\ &\quad - b_j(t)\sum_{l=1}^{T} I\{o_l = t\mid\lambda\}\alpha_l(j)\beta_l(j)\\ &\quad - \sum_{z=1,z\neq t}^{M} b_j(t)\sum_{l=1}^{T} I\{o_l = z\mid\lambda\}\alpha_l(j)\beta_l(j) \qquad \text{(A.36)}\end{aligned}$$

Τοποθετώντας μαζί τους δύο τελευταίους όρους της δεξιάς πλευράς της (A.36) λαμβάνουμε

$$\frac{\partial P(O\mid\lambda)}{\partial r_{jt}} = \sum_{l=1}^{T} I\{o_l = t\mid\lambda\}\alpha_l(j)\beta_l(j) - \sum_{z=1}^{M} b_j(t)\sum_{l=1}^{T} I\{o_l = z\mid\lambda\}\alpha_l(j)\beta_l(j) \qquad \text{(A.37)}$$

εναλλάσσοντας την σειρά των αθροίσεων στον δεύτερο όρο της (A.37) λαμβάνουμε

$$\sum_{z=1}^{M} b_j(t)\sum_{l=1}^{T} I\{o_l = z\mid\lambda\}\alpha_l(j)\beta_l(j) = b_j(t)\sum_{l=1}^{T}\alpha_l(j)\beta_l(j)\sum_{z=1}^{M} I\{o_l = z\mid\lambda\} \qquad \text{(A.38)}$$

κάθε σύμβολο o_l μιας ακολουθίας παρατηρήσεων O, λαμβάνει μια τιμή από το V (μια από τις M δυνατές τιμές v_m), οπότε άθροιση επί της συνάρτησης δείκτη δίνει

$$\sum_{z=1}^{M} I\{o_l = z\mid\lambda\} = 1 \qquad \text{(A.39)}$$

από την (A.38) με χρήση της (A.39) λαμβάνουμε

$$b_j(t)\sum_{l=1}^{T}\alpha_l(j)\beta_l(j)\sum_{z=1}^{M} I\{o_l = z\mid\lambda\} = b_j(t)\sum_{l=1}^{T}\alpha_l(j)\beta_l(j) \qquad \text{(A.40)}$$

εισαγωγή της ανωτέρω έκφρασης (A.40) στην (A.37) δίνει

$$\frac{\partial P(O \mid \lambda)}{\partial r_{jt}} = \sum_{l=1}^{T} I\{o_l = t \mid \lambda\}\alpha_l(j)\beta_l(j) - b_j(t)\sum_{l=1}^{T}\alpha_l(j)\beta_l(j)$$

$$= \sum_{l=1}^{T}\left[I\{o_l = t \mid \lambda\}\alpha_l(j)\beta_l(j) - b_j(t)\alpha_l(j)\beta_l(j)\right]$$

οπότε και αποδεικνύεται η (A.35).

Λήμμα 5:

$$\frac{\partial P(O \mid \lambda)}{\partial \pi_r} = b_r(o_1)\beta_1(\underline{r}) \qquad \text{(A.41)}$$

Απόδειξη: Από τον τύπο αποτίμησης (A.11) για t = 1 λαμβάνουμε

$$P(O \mid \lambda) = \sum_{j=1}^{N}\alpha_1(j)\beta_1(j) \qquad \text{(A.42)}$$

διαφορίζουμε μερικώς (A.42) ως προς το π_r, οπότε

$$\frac{\partial P(O \mid \lambda)}{\partial \pi_r} = \sum_{j=1}^{N}\frac{\partial \alpha_1(j)}{\partial \pi_r}\beta_1(j). \quad \text{(A.43)}$$

Συνεπώς, απαιτείται μόνο ο υπολογισμός του όρου στη δεξιά πλευρά. Αυτό επιτυγχάνεται μια ακόμα φορά με χρήση της πρόσθιας αναδρομής (A.7)

$$\frac{\partial \alpha_1(j)}{\partial \pi_r} = \frac{\partial \pi_j b_j(o_1)}{\partial \pi_r} = b_j(o_1)\frac{\partial \pi_j}{\partial \pi_r} \qquad \text{(A.44)}$$

Διαφορίζοντας, λαμβάνουμε

$$\frac{\partial \pi_r}{\partial \pi_r} = 1,$$

$$\frac{\partial \pi_j}{\partial \pi_r} = 0, \quad j \neq r$$

με χρήση των μερικών παραγώγων στον δεύτερο όρο της (A.43) λαμβάνεται

$$\frac{\partial P(O \mid \lambda)}{\partial \pi_r} = \sum_{j=1}^{N} b_j(o_1)\frac{\partial \pi_j}{\partial \pi_r}\beta_1(j) = b_r(o_1)\beta_1(r)$$

άρα αποδεικνύεται η (A.41).

Λήμμα 6:

$$\frac{\partial P(O \mid \lambda)}{\partial u_j} = \pi_j \left[b_j(o_1)\beta_1(j) - P(O \mid \lambda) \right] \qquad \text{(A.45)}$$

Απόδειξη: Χρησιμοποιώντας τον κανόνα τη αλυσίδας στην αριστερή πλευρά της (A.45), λαμβάνουμε

$$\frac{\partial P(O \mid \lambda)}{\partial u_j} = \sum_{z=1}^{N} \frac{\partial P(O \mid \lambda)}{\partial \pi_z} \frac{\partial \pi_z}{\partial u_j} = \frac{\partial P(O \mid \lambda)}{\partial \pi_j} \frac{\partial \pi_j}{\partial u_j} + \sum_{z=1, z \neq j}^{N} \frac{\partial P(O \mid \lambda)}{\partial \pi_z} \frac{\partial \pi_z}{\partial w_j}$$

εφαρμόζοντας τις (A.5), (A.6) και (A.41) στην δεξιά πλευρά της ανωτέρω έκφρασης λαμβάνουμε

$$\begin{aligned} \frac{\partial P(O \mid \lambda)}{\partial u_j} &= \pi_j (1 - \pi_j) b_j(o_1)\beta_1(j) - \sum_{z=1, z \neq j}^{N} \pi_z \pi_j b_z(o_1)\beta_1(z) \\ &= \pi_j b_j(o_1)\beta_1(j) \\ &- \pi_j \left[\pi_j b_j(o_1)\beta_1(j) \right] \\ &- \pi_j \sum_{z=1, z \neq j}^{N} \pi_z b_z(o_1)\beta_1(z). \qquad \text{(A.46)} \end{aligned}$$

Τοποθετώντας τους δύο τελευταίους όρους της δεξιάς πλευράς της (A.46) μαζί λαμβάνουμε

$$\frac{\partial P(O \mid \lambda)}{\partial u_j} = \pi_j \left[b_j(o_1)\beta_1(j) - \sum_{z=1}^{N} \pi_z b_z(o_1)\beta_1(z) \right] \qquad \text{(A.47)}$$

χρησιμοποιώντας την οπίσθια αναδρομή της (A.7), λαμβάνουμε

$$\sum_{z=1}^{N} \pi_z b_z(o_1)\beta_1(z) = \sum_{z=1}^{N} \alpha_1(z)\beta_1(z)$$

σύμφωνα με την (A.11) η ανωτέρω έκφραση δίνει

$$\sum_{z=1}^{N} \alpha_1(z)\beta_1(z) = P(O \mid \lambda) \qquad \text{(A.48)}$$

με χρήση της (A.48) στην (A.47) λαμβάνουμε

$$\frac{\partial P(O \mid \lambda)}{\partial u_j} = \pi_j \left[b_j(o_1)\beta_1(j) - P(O \mid \lambda) \right]$$

άρα, αποδεικνύεται η (A.45).

Παράρτημα Β

Η βασική διαδικασία μετατροπής κλίμακας που χρησιμοποιείται είναι να πολλαπλασιαστουν οι $\alpha_i(t)$ με έναν συντελεστή στάθμισης ανεξάρτητο του i (δηλαδή εξαρτώμενου μόνο από τον χρόνο t), με στόχο να διατηρηθούν τα υπό κλίμακα $\alpha_i(t)$ εντός της δυνατότητας ακρίβειας του υπολογιστή για $1 \leq t \leq T$. Μια παρόμοια διαδικασία μετατροπής κλίμακας πραγματοποιείται στους συντελεστές $\beta_i(t)$ (αφού και αυτοί τείνουν στο μηδέν με εκθετική ταχύτητα) και στη συνέχεια, στο τέλος του υπολογισμού, οι συντελεστές κλίμακας αναιρούνται.

Η υπό κλίμακα πρόσθια μεταβλητή and η υπό κλίμακα οπίσθια μεταβλητή ορίζονται ως

$$\hat{\alpha}_t(i) = \frac{P(o_1 o_2 \ldots o_t, q_t = s_i \mid \lambda)}{P(o_1 o_2 \ldots o_t \mid \lambda)} \qquad \text{(B.1)}$$

$$\hat{\beta}_t(i) = \frac{P(o_{t+1} o_{t+2} \ldots o_T \mid q_t = s_i, \lambda)}{P(o_{t+1} o_{t+2} \ldots o_T \mid o_1 o_2 \ldots o_t, \lambda)} \qquad \text{(B.2)}$$

Μια αναδρομική λύση για την υπό κλίμακα πρόσθια μεταβλητή $\hat{\alpha}_t(i)$ είναι η ακόλουθη:

1] Αρχικοποίηση

$$\tilde{\alpha}_1(i) = \pi_i b_i(o_1)$$

$$\hat{\alpha}_1(i) = sc_1 \tilde{\alpha}_1(i), \qquad 1 \leq i \leq N$$

2] Αναδρομή:

$$\tilde{\alpha}_t(i) = \left[\sum_{j=1}^{N} \hat{\alpha}_{t-1}(j) a_{ji} \right] b_i(o_t)$$

$$\hat{\alpha}_t(i) = sc_t \tilde{\alpha}_t(i), \qquad 1 \leq i \leq N,\ 2 \leq t \leq T$$

όπου ο συντελεστής κλίμακας sc_t ορίζεται ως

$$sc_t = \frac{1}{\sum_{i=1}^{N} \tilde{\alpha}_t(i)} \qquad \text{(B.3)}$$

άρα, επαγωγικά λαμβάνουμε

$$\hat{\alpha}_t(i) = \left[\prod_{\tau=1}^{t} sc_\tau \right] \alpha_t(i) \qquad \text{(B.4)}$$

όπου $\alpha_t(i)$ είναι η πρόσθια μεταβλητή που μπορεί να υπολογιστεί σύμφωνα με τον forward-backward αλγόριθμο.

Στη συνέχεια, υπολογίζουμε την υπό κλίμακα οπίσθια μεταβλητή $\hat{\beta}_t(i)$ από την οπίσθια αναδρομή. Η μόνη διαφορά είναι ότι χρησιμοποιούμε τους ίδιοις συντελεστές

κλίμακας για κάθε t. Μια αναδρομική λύση γι την η υπό κλίμακα οπίσθια μεταβλητή $\hat{\beta}_t(i)$ είναι η ακόλουθη:

1] Αρχικοποίηση:

$$\tilde{\beta}_T(i) = 1$$

$$\hat{\beta}_T(i) = sc_T \tilde{\beta}_T(i), \qquad 1 \le i \le N$$

2] Αναδρομή:

$$\tilde{\beta}_t(i) = \sum_{j=1}^{N} a_{ij} b_i(o_{t+1}) \hat{\beta}_{t+1}(j)$$

$$\hat{\beta}_t(i) = sc_t \tilde{\beta}_t(i), \qquad 1 \le i \le N,\ t = T\text{-}1,\ T\text{-}2,\ \ldots,\ 1$$

Αφού κάθε συντελεστής κλίμακας επαναφέρει αποτελεσματικά το μέγεθος των πρόσθιων μεταβλητών και οι οπίσθιες μεταβλητές είναι συγκρίσιμες με αυτούς, ένας αποτελεσματικός τρόπος να διατηρηθεί ο υπολογισμός σε λογικά όρια είναι να χρησιμοποιηθεί ο ίδιος συντελεστής κλίμακας.

Όμοια, με επαγωγή λαμβάνουμε

$$\hat{\beta}_t(i) = \left[\prod_{\tau=t}^{T} sc_\tau \right] \beta_t(i) \qquad \text{(B.5)}$$

όπου $\beta_t(i)$ ειναι η οπίσθια μεταβλητή που μπορεί να υπολογιστεί σύμφωνα με τον forward-backward αλγόριθμο.

Είναι σημαντικό να τνιστεί ότι ο όρος

$$\prod_{\tau=1}^{t} sc_\tau \prod_{\tau=t+1}^{T} sc_\tau = \prod_{\tau=1}^{T} sc_\tau = SC_T \qquad \text{(B.6)}$$

είναι ανεξάρτητος του t και εξαρτάται μόνο από το μήκος της ακολουθίας παρατηρήσεων T.

Η μόνη πραγματική αλλαγή που συμβαίνει εξαιτίας της κλίμακας αφορά τον υπολογισμό της πιθανότητας $P(O \mid \lambda)$. Αρχικά, παρατηρούμε πως

$$\sum_{j=1}^{N} \hat{\alpha}_T(j) = \sum_{j=1}^{N} sc_T \tilde{\alpha}_T(j) = \sum_{j=1}^{N} \frac{1}{\sum_{i=1}^{N} \tilde{\alpha}_T(i)} \tilde{\alpha}_T(j) = \frac{\sum_{j=1}^{N} \tilde{\alpha}_T(j)}{\sum_{i=1}^{N} \tilde{\alpha}_T(i)} = 1$$

επίσης, με χρήση της (B.4) έχουμε

$$\sum_{j=1}^{N} \hat{\alpha}_T(j) = \sum_{j=1}^{N} \left(\prod_{\tau=1}^{T} sc_\tau \right) \alpha_T(j) = SC_T \sum_{j=1}^{N} \alpha_T(j) = SC_T P(O \mid \lambda)$$

όπου η τελευταία ισότητα προκύπτει από τον ορισμό της πρόσθιας μεταβλητής. Κατά συνέπεια,, η πιθανότητα ως προς τις υπό κλίμακα πρόσθιες/οπίσθιες μεταβλητές μπορεί να

υπολογιστεί με χρήση των ακόλουθων τύπων

$$P(O \mid \lambda) = \frac{1}{\prod_{t=1}^{T} sc_t} = \frac{1}{SC_T} \qquad \text{(B.7)}$$

$$\log P(O \mid \lambda) = -\log\left[\prod_{t=1}^{T} sc_t\right] = -\sum_{t=1}^{T} \log(sc_t) \qquad \text{(B.8)}$$

άρα, μπορεί να υπολογιστεί η λογαριθμική πιθανότητα αντί της απλής πιθανότητας, αφού θα ήταν εκτός των ορίων ακρίβειας της μηχανής διαφορετικά.

Είναι προφανές ότι η διαδικασία αλλαγής κλίμακας δεν χρειάζεται να εφαρμοστεί σε κάθε χρονική στιγμή *t*, αλλά μπορεί να πραγματοποιηθεί όποτε επιθυμείται ή απαιτείται. Αν δεν πραγματοποιηθεί η αλλαγή κλίμακας κάποια στιγμή *t*, οι συντελεστές κλίμακας τίθονται 1 τη στιγμή αυτή και όλες οι συνθήκες που συζητήθηκαν ανωτέρω ικανοποιούνται.

Λήμμα 7:

$$\sum_{l=1}^{T-1}\left[\hat{\alpha}_l(i) b_j(o_{l+1}) \hat{\beta}_{l+1}(j) - \left[\sum_{k=1}^{N} \hat{\alpha}_{l-1}(k) a_{ki}\right] b_i(o_l) \hat{\beta}_l(i)\right]$$
$$= \frac{1}{P(O|\lambda)} \sum_{l=1}^{T-1}\left[\alpha_l(i) b_j(o_{l+1}) \beta_{l+1}(j) - \alpha_l(i) \beta_l(i)\right] \qquad \text{(B.9)}$$

Απόδειξη: με χρήση των (B.4) και (B.5) στην αριστερή πλευρά της (B.9) λαμβάνουμε

$$\sum_{l=1}^{T-1}\left[\hat{\alpha}_l(i) b_j(o_{l+1}) \hat{\beta}_{l+1}(j) - \left[\sum_{k=1}^{N} \hat{\alpha}_{l-1}(k) a_{ki}\right] b_i(o_l) \hat{\beta}_l(i)\right]$$
$$= \sum_{l=1}^{T-1}\left\{\left(\lambda_{\tau=1}^{l} sc_\tau\right) \alpha_l(i) b_j(o_{l+1}) \left(\lambda_{\tau=l+1}^{T} sc_\tau\right) \beta_{l+1}(j)\right.$$
$$\left. - \left[\sum_{k=1}^{N}\left(\lambda_{\tau=1}^{l-1} sc_\tau\right) \alpha_{l-1}(k) a_{ki}\right] b_i(o_l) \left(\lambda_{\tau=l}^{T} sc_\tau\right) \beta_l(i)\right\}$$
$$= \sum_{l=1}^{T-1}\left\{\left(\lambda_{\tau=1}^{T} sc_\tau\right) \alpha_l(i) b_j(o_{l+1}) \beta_{l+1}(j) - \left(\lambda_{\tau=1}^{T} sc_\tau\right)\left[\sum_{k=1}^{N} \alpha_{l-1}(k) a_{ki}\right] b_i(o_l) \beta_l(i)\right\}$$
$$\text{(B.10)}$$

χρησιμοποιώντας την (B.6) λαμβάνουμε

$$\sum_{l=1}^{T-1}\left\{\left(\lambda_{\tau=1}^{T} sc_\tau\right)\alpha_l(i)b_j(o_{l+1})\beta_{l+1}(j)-\left(\lambda_{\tau=1}^{T} sc_\tau\right)\left[\sum_{k=1}^{N}\alpha_{l-1}(k)a_{ki}\right]b_i(o_l)\beta_l(i)\right\}$$

$$= SC_T\sum_{l=1}^{T-1}\left[\alpha_l(i)b_j(o_{l+1})\beta_{l+1}(j)-\left[\sum_{k=1}^{N}\alpha_{l-1}(k)a_{ki}\right]b_i(o_l)\beta_l(i)\right] \qquad \text{(B.11)}$$

αφού το SC_T είναι ανεξάρτητο του l. Χρησιμοποιώντας την πρόσθια αναδρομή του forward-backward αλγορίθμου στην δεξιά πλευρά της (B.11) λαμβάνουμε

$$SC_T\sum_{l=1}^{T-1}\left[\alpha_l(i)b_j(o_{l+1})\beta_{l+1}(j)-\left[\sum_{k=1}^{N}\alpha_{l-1}(k)a_{ki}\right]b_i(o_l)\beta_l(i)\right]$$

$$= SC_T\sum_{l=1}^{T-1}\left[\alpha_l(i)b_j(o_{l+1})\beta_{l+1}(j)-\alpha_l(i)\beta_l(i)\right]$$

$$= \frac{1}{P(O\mid\lambda)}\sum_{l=1}^{T-1}\left[\alpha_l(i)b_j(o_{l+1})\beta_{l+1}(j)-\alpha_l(i)\beta_l(i)\right] \qquad \text{(B.12)}$$

όπου εισήχθη η (B.7). Χρησιμοποιώντας τις (B.10) – (B.12) βλέπουμε ότι η (B.9) αποδεικνύεται.

Λήμμα 8:

$$\sum_{l=1}^{T}\left[I\{o_l=t\mid\lambda\}\left[\sum_{i=1}^{N}\hat{\alpha}_{l-1}(i)a_{ij}\right]b_j(o_l)\hat{\beta}_l(j)-b_j(t)\left[\sum_{i=1}^{N}\hat{\alpha}_{l-1}(i)a_{ij}\right]b_j(o_l)\hat{\beta}_l(j)\right]$$

$$= \frac{1}{P(O\mid\lambda)}\sum_{l=1}^{T}\left[I\{o_l=t\mid\lambda\}\alpha_l(j)\beta_l(j)-b_j(t)\alpha_l(j)\beta_l(j)\right] \qquad \text{(B.13)}$$

Απόδειξη: με χρήση των (B.4) και (B.5) στην αριστερή πλευρά της (B.13) λαμβάνουμε

$$\sum_{l=1}^{T}\left[I\{o_l=t\mid\lambda\}\left[\sum_{i=1}^{N}\hat{\alpha}_{l-1}(i)a_{ij}\right]b_j(o_l)\hat{\beta}_l(j)-b_j(t)\left[\sum_{i=1}^{N}\hat{\alpha}_{l-1}(i)a_{ij}\right]b_j(o_l)\hat{\beta}_l(j)\right]$$

$$= \sum_{l=1}^{T}\left\{I\{o_l=t\mid\lambda\}\left[\sum_{i=1}^{N}\left(\prod_{\tau=1}^{l-1}sc_\tau\right)\alpha_{l-1}(i)a_{ij}\right]b_j(o_l)\left(\prod_{\tau=l}^{T}sc_\tau\right)\beta_l(j)\right.$$

$$\left. - b_j(t)\left[\sum_{i=1}^{N}\left(\prod_{\tau=1}^{l-1}sc_\tau\right)\alpha_{l-1}(i)a_{ij}\right]b_j(o_l)\left(\prod_{\tau=l}^{T}sc_\tau\right)\beta_l(j)\right\}$$

$$= \sum_{l=1}^{T}\left\{I\{o_l=t\mid\lambda\}\left(\prod_{\tau=1}^{T}sc_\tau\right)\left[\sum_{i=1}^{N}\alpha_{l-1}(i)a_{ij}\right]b_j(o_l)\beta_l(j)\right.$$

$$- b_j(t)\left(\prod_{\tau=1}^{T} sc_\tau\right)\left[\sum_{i=1}^{N}\alpha_{l-1}(i)a_{ij}\right]b_j(o_l)\beta_l(j)\Bigg\} \qquad \text{(B.14)}$$

χρησιμοποιώντας την (B.6) λαμβάνουμε

$$\sum_{l=1}^{T}\left\{I\{o_l=t|\lambda\}\left(\prod_{\tau=1}^{T} sc_\tau\right)\left[\sum_{i=1}^{N}\alpha_{l-1}(i)a_{ij}\right]b_j(o_l)\beta_l(j)-b_j(t)\left(\prod_{\tau=1}^{T} sc_\tau\right)\left[\sum_{i=1}^{N}\alpha_{l-1}(i)a_{ij}\right]b_j(o_l)\beta_l(j)\right\}$$

$$=$$

$$SC_T\sum_{l=1}^{T}\left\{I\{o_l=t|\lambda\}\left[\sum_{i=1}^{N}\alpha_{l-1}(i)a_{ij}\right]b_j(o_l)\beta_l(j)-b_j(t)\left[\sum_{i=1}^{N}\alpha_{l-1}(i)a_{ij}\right]b_j(o_l)\beta_l(j)\right\} \qquad \text{(B.15)}$$

αφού το SC_T is ανεξάρτητο του l. Χρησιμοποιώντας την πρόσθια αναδρομή του forward-backward αλγορίθμου στην δεξιά πλευρά της (B.15) λαμβάνουμε

$$SC_T\sum_{l=1}^{T}\left\{I\{o_l=t|\lambda\}\left[\sum_{i=1}^{N}\alpha_{l-1}(i)a_{ij}\right]b_j(o_l)\beta_l(j)-b_j(t)\left[\sum_{i=1}^{N}\alpha_{l-1}(i)a_{ij}\right]b_j(o_l)\beta_l(j)\right\}$$

$$= SC_T\sum_{l=1}^{T}\left\{I\{o_l=t|\lambda\}\alpha_l(j)\beta_l(j)-b_j(t)\alpha_l(j)\beta_l(j)\right\}$$

$$= \frac{1}{P(O|\lambda)}\sum_{l=1}^{T}\left\{I\{o_l=t|\lambda\}\alpha_l(j)\beta_l(j)-b_j(t)\alpha_l(j)\beta_l(j)\right\} \qquad \text{(B.16)}$$

όπου εισήχθη η (B.7). Χρησιμοποιώντας τις (B.14)–(B.16) αποδεικνύεται η (B.13).

Λήμμα 9:

$$b_j(o_1)\hat{\beta}_1(j) = \frac{b_j(o_1)\beta_1(j)}{P(O|\lambda)} \qquad \text{(B.17)}$$

Απόδειξη: Θέτοντας την (B.5) στην αριστερή πλευρά της (B.17) λαμβάνουμε

$$b_j(o_1)\hat{\beta}_1(j) = b_j(o_1)\left(\prod_{\tau=1}^{T} sc_\tau\right)\beta_1(j) \qquad \text{(B.18)}$$

με χρήση της (B.6) λαμβάνουμε

$$b_j(o_1)\left(\prod_{\tau=1}^{T} sc_\tau\right)\beta_1(j) = SC_T b_j(o_1)\beta_1(j) \qquad \text{(B.19)}$$

εισάγωντας την (B.7) στην δεξιά πλευρά της προηγούμενης έκφρασης λαμβάνουμε

$$SC_T b_j(o_1)\beta_1(j) = \frac{b_j(o_1)\beta_1(j)}{P(O|\lambda)} \qquad \text{(B.20)}$$

και με χρήση των (B.18) – (B.20) αποδεικνύεται η (B.17).