Αντίστοιχη Γονιδιακής Έκφρασης με Βαγγέλη Λίχτενη

(36)
ΜΠΛΕ

Η ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ ΕΞΕΙΔΙΚΕΥΣΗΣ

υποβλήθηκε στην

Ομοτιμότητα της Γενικής Συνέλευσης Ειδικής Σύνθεσης

του Τμήματος Πληροφορικής Εξεταστική Επιτροπή

από τον

Γεώργιο Ρήγα

ως μέρος των Υποχρεώσεων για τη λήψη του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

ΜΕ ΕΞΕΙΔΙΚΕΥΣΗ ΣΤΙΣ ΤΕΧΝΟΛΟΓΙΕΣ ΕΦΑΡΜΟΓΩΝ

Μάρτιος 20[illegible]

Ανάλυση Γονιδιακής Έκφρασης με Bayesian Δίκτυα

Η ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ ΕΞΕΙΔΙΚΕΥΣΗΣ

υποβάλλεται στην

ορισθείσα από την Γενική Συνέλευση Ειδικής Σύνθεσης

του Τμήματος Πληροφορικής Εξεταστική Επιτροπή

από τον

Γεώργιο Ρήγα

ως μέρος των Υποχρεώσεων για τη λήψη του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

ΜΕ ΕΞΕΙΔΙΚΕΥΣΗ ΣΤΙΣ ΤΕΧΝΟΛΟΓΙΕΣ-ΕΦΑΡΜΟΓΕΣ

Μάρτιος 2006

ΕΥΧΑΡΙΣΤΙΕΣ

Για την εκπόνηση αυτής της διατριβής οφείλω να ευχαριστήσω κυρίως τον επιβλέποντα καθηγητή κ.κ. Δ. Ι. Φωτιάδη, για την πολύτιμη βοήθεια του, την εμπιστοσύνη που μου έδειξε και τον χρόνο που αφιέρωσε για μένα. Επίσης μεγάλο ευχαριστώ οφείλω στους φίλους συναδέλφους μου Α. Φωτίου, Ι. Καραμήτσιο, Κ. Στεφανίδη, Ε. Οικονόμου, Μ. Τσίπουρα, Κ. Βόγκλη και Σ. Πέτσιο για την αμέριστη συμπαράσταση τους, όσο και στην Ε. Μπαϊκούση για την μεγάλη της βοήθεια στην συγγραφή της διατριβής.

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΥΡΕΤΗΡΙΟ ΣΧΗΜΑΤΩΝ

Ευρετηριο Πινακων

ΕΠΕΞΗΓΗΣΕΙΣ ΣΥΜΒΟΛΙΣΜΩΝ

D	Σύνολο δεδομένων
N	Αριθμός δεδομένων στο σύνολο δεδομένων π.χ $\|D\|$
X,Y,Z,..	Μονοδιάστατες μεταβλητές
x,y,z,..	Τιμές για τις αντίστοιχες μεταβλητές
S,**T**, ...	Σύνολα μεταβλητών
s,**t**,...	Τιμές για τα αντίστοιχα σύνολα μεταβλητών
$\mathcal{U}$	Χώρος που περιέχει l μεταβλητές από ένα πεδίο : $X_1, ..., X_n$
n	Αριθμός μεταβλητών, π.χ $\|\mathcal{U}\|$
$\mathcal{G}$	Ακυκλικό μη κατευθυνόμενο γράφημα (DAG)
$\mathcal{B}$	Bayesian δίκτυο.
M	Δομή ενός Bayesian δικτύου.
$\mathfrak{M}$	Το σύνολο των πιθανών δομών ενός Bayesian δικτύου.
Pa_i	Το σύνολο πατέρων της μεταβλητής X_i
pa_{ij}	Ανάθεση τιμών j για τους πατέρες της μεταβλητής X_i
θ	Παράμετροι
Θ	Σύνολο παραμέτρων
α_i	Υπερπαράμετροι για την πολυνυμική κατανομή
r_i	Αριθμός των πιθανών τιμών για την διακριτή μεταβλητή X_i
q_i	Αριθμός των πιθανών συνδυασμών τιμών των πατέρων της μεταβλητής X_i
N_i	Ο αριθμός των εμφανίσεων στα δεδομένα της τιμής i για πολυνυμικές κατανομές.
ϕ, ψ	Δυναμικά κλικών
$\sum_i$	Όταν παραλείπεται το άνω όριο εννοείται η άθροιση πάνω σε όλες τις πιθανές τιμές.

Περιληψη

Γεώργιος Ρήγας του Αντωνίου και της Βασιλικής. MSc, Τμήμα Πληροφορικής, Πανεπιστήμιο Ιωαννίνων, Απρίλιος, 2006. Ανάλυση Γονιδιακής Έκφρασης με Bayesian Δίκτυα. Επιβλέποντας: Δημήτριος Ι. Φωτιάδης.

Τα τελευταία δέκα χρόνια έχει γίνει πολύ έρευνα στα Bayesian δίκτυα τα οποία έχουν βρει εφαρμογή σε πολλούς τομείς. Ένας τομέας που έχει μελετηθεί αρκετά είναι η εφαρμογή τους για την εύρεση των ρυθμιστικών δικτύων γονιδίων. Παρόλα αυτά πολλές προκλήσεις παραμένουν ανοιχτές. Το πρώτο πρόβλημα που προσπαθούμε να αντιμετωπίσουμε είναι ο χειρισμός ελλιπών τιμών στα δεδομένα στην διαδικασία μάθησης των Bayesian δικτύων. Υιοθετούμε την προσέγγιση που πρότεινε ο N. Friedman με τον Structural EM όπου ο χειρισμός των ελλιπών τιμών γίνεται στην διαδικασία μάθησης της δομής του Bayesian δικτύου. Το δεύτερο πρόβλημα που καλούμαστε να αντιμετωπίσουμε είναι ότι ο χρόνος εκπαίδευσης ενός δικτύου με μεγάλο αριθμό μεταβλητών και μάλιστα με ελλιπής τιμές είναι πολύ μεγάλος. Προσπαθούμε να δώσουμε μια λύση στο πρόβλημα με μία παράλληλη υλοποίηση της όλης διαδικασίας. Τα πειράματα είναι αρκετά ικανοποιητικά σε σχέση με την απόδοση της παραλληλίας και σχετικά με την εφαρμογή των Bayesian δικτύων σε προβλήματα με πολλές μεταβλητές.

Extended Abstract in English

George, Rigas. MSc, Computer Science Department, University of Ioannina, Greece, April, 2006. Reconstruction of Genetic networks using bayesian networks. Supervisor: Dimitrios I. Fotiadis.

In the last decade many researchers have focused on Bayesian networks that have been applied in various fields. One of these fields is the application of Bayesian networks in order to learn the structure of the regulatory networks. However, there are still a number of challenges that remain open. The first challenge that we are trying to resolve is how to handle missing values in the procedure of learning Bayesian networks. We adopt the approach of Structural EM proposed by N. Friedman, where the way of handling missing values is incorporated in the learning of the Bayesian network's structure. The second challenge trying to resolve is concerning the fact that the learning time of a network containing a great number of variables with missing values is enormous. The solution we propose is by a parallel implementation of the whole procedure. The results are quite satisfactory when concerning the parallel implementation. Therefore, we are capable of applying Bayesian networks in problems that contain a great number of variables.

Κεφαλαιο 1

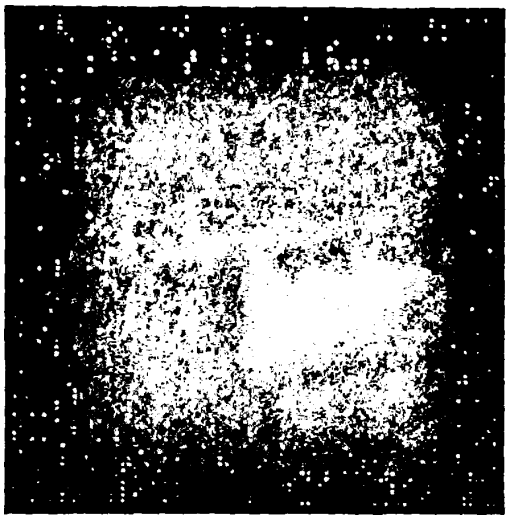
Εισαγωγη

1.1 Περιγραφή Προβλήματος

Η κλασική προσέγγιση έρευνας στην μοριακή βιολογία υπήρξε η εξέταση και συλλογή δεδομένων για μια μεμονωμένη οντότητα, επικεντρωνόταν στην μελέτη ενός μόνο γονιδίου, μίας μόνο πρωτείνης ή μιας μόνο αντίδρασης ξεχωριστά. Αυτή η προσέγγιση έχει δώσει αξιοσημείωτα αποτελέσματα, επιτρέποντάς μας ακριβή βιολογικά μοντέλα.

Ωστόσο, με το πέρασμα στην «Εποχή των Γονιδίων» έχει προκύψει μία τελείως διαφορετική κλάση δεδομένων. Στο παρελθόν, η ανάλυση αυτού του μεγάλου όγκου δεδομένων, αποτελούνταν από απλές περιγραφές του πόσα γονίδια ήταν μέχρι εκείνη την στιγμή άγνωστα, ποια γονίδια είναι υπό ή υπερ εκφρασμένα υπό ορισμένες συνθήκες, κλπ. Αυτά τα δεδομένα είναι φυσικά χρήσιμα σε ερευνητές που εστιάζονται σε μερικά γονίδια. Αλλά δεν μπορούμε να περιμένουμε να κατασκευάσουμε ένα πλήρες βιολογικό μοντέλο, για παράδειγμα του γονιδιώματος της ζύμης που έχει περίπου 6000 γονίδια αναλύοντας το κάθε γονίδιο ξεχωριστά.

Έτσι, προκύπτει η ανάγκη για μεθόδους που μπορούν να χειριστούν δεδομένα μαζικά και που μπορούν να αναλύσουν μεγάλα συστήματα σε κάποιο ενδιάμεσο στάδιο, χωρίς να φτάνουν μέχρι τις βιοχημικές αντιδράσεις. Τα τελευταία χρόνια, μια τέτοιου είδους ανάλυση



Σχήμα 1.1: Το γονιδίωμα της ζύμης σε μία μικροσυστοιχία

θα μπορούσε να βοηθήσει στην καθοδήγηση των κλασικών προσεγγίσεων στην φαρμακολογία και στην βιοχημεία ως χάρτης για τα γονίδια που χρήζουν περαιτέρω μελέτης, ανάμεσα στα χιλιάδες που έχουν ανακαλυφθεί. Ιδανικά, ένα σύστημα με ικανοποιητικές ιδιότητες πρόβλεψης και ανάλυσης σε ένα ενδιάμεσο στάδιο θα μπορούσε να εξαλείψει την ανάγκη για πλήρη κατανόηση του όλου συστήματος σε βιοχημικό επίπεδο. Η ανάλυση της γονιδιακής έκφρασης βασίζεται στην υπόθεση ότι η πληροφορία για την λειτουργική κατάσταση ενός οργανισμού καθορίζεται σε μεγάλο βαθμό από την πληροφορία στην έκφραση των γονιδίων.

Ένα **ρυθμιστικό δίκτυο** γονιδίων (ή **γενετικό δίκτυο**) είναι μια συλλογή τμημάτων του DNA σε ένα κύτταρο που αλληλεπιδρούν μεταξύ τους και με άλλες ουσίες στο κύτταρο, ελέγχοντας με αυτόν τον τρόπο τους ρυθμούς με τους οποίους κάθε γονίδιο μεταγράφεται σε mRNA.

Τα ρυθμιστικά δίκτυα γονιδίων μόλις πρόσφατα έχουν αρχίσει να κατανοούνται και η ανακάλυψη των λειτουργιών κάθε γονιδίου αποτελεί το επόμενο βήμα έρευνας στην βιολογία, ώστε να μπορέσει να μοντελοποιηθεί η συμπεριφορά του κυττάρου.

Για να αποκτήσουμε μια πλήρη εικόνα από τα δεδομένα της γονιδιακής έκφρασης, είναι απαραίτητο κάθε γονίδιο να παρατηρηθεί υπό αρκετά διαφορετικές συνθήκες και προτιμότερο σε μορφή χρονο-ακολουθιών έκφρασης [1]. Τέτοια δεδομένα μπορούν να αναλυθούν χρησιμοποιώντας μία μεγάλη ποικιλία μεθόδων όπως θα δούμε στην συνέχεια.

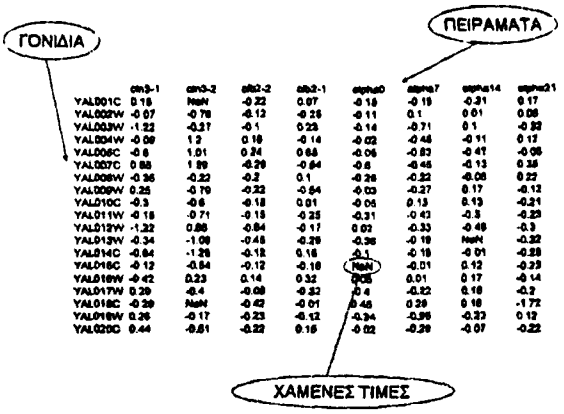
1.2 Περιγραφή των δεδομένων

Το εργαλείο που μας δίνει τα δεδομένα είναι οι μικροσυστοιχίες DNA (DNA microarrays ή DNA chips). Η μικροσυστοιχία είναι μια μικρή πλάκα και έχει στην επιφάνειά της πάρα πολλές οπές. Μέσα σε αυτές τις οπές τοποθετείται καταρχήν το DNA του γονιδίου που θέλουμε να ανιχνεύσουμε. Συγκεκριμένα τοποθετείται μονόκλωνο DNA του φυσιολογικού γονιδίου (DNA που γνωρίζουμε ότι δεν φέρει μεταλλάξεις) το οποίο προσδένεται στην επιφάνεια του chip με κάποιες τεχνικές (το DNA κανονικά είναι δίκλωνο μόριο δηλαδή αποτελείται από δύο συμπληρωματικές αλυσίδες). Στις οπές του chip τοποθετούμε από πάνω, αυτόματα με ειδικά μηχανήματα DNA (του συγκεκριμένου γονιδίου σε πολλά μικρά κομματάκια) του ασθενή (αυτού που θέλουμε να δούμε αν φέρει μεταλλάξεις στο συγκεκριμένο γονίδιο) το οποίο έχουμε κάνει και αυτό μονόκλωνο και το έχουμε χρωματίσει πράσινο. Εκτός από του ασθενή τοποθετούμε και DNA φυσιολογικού ατόμου το οποίο έχουμε χρωματίσει κόκκινο. Στην συνέχεια αφήνουμε τα DNA να υβριδοποιηθούν (να ενωθούν οι βάσεις μεταξύ τους με συγκεκριμένο τρόπο) σε κατάλληλες συνθήκες (να κολλήσουν μεταξύ τους αφού όπως αναφέραμε οι αλυσίδες είναι συμπληρωματικές). Σε αυτήν την φάση το αποτέλεσμα δεν είναι ορατό. Για να δούμε το αποτέλεσμα βάζουμε το chip σε ένα μαύρο κουτί και στην συνέχεια το σαρώνουμε πρώτα με ένα λέϊζερ πράσινου χρώματος και στην συνέχεια με ένα κόκκινου χρώματος και αποθηκεύουμε τις εικόνες. Στην συνέχεια κάνουμε σύντηξη των δύο εικόνων που προκύπτουν. Ανάλογα με το χρώμα που θα προκύψει από την υβριδοποίηση των κομματιών DNA μπορούμε να ανιχνεύσουμε πιθανές μεταλλάξεις στο γονίδιο που μελετάμε. Όλη η παραπάνω διαδικασία γίνεται αυτόματα και το μόνο που χρειάζεται να κάνουμε είναι να ετοιμάσουμε τα δείγματα και να επεξεργαστούμε τα αποτελέσματα. Η τεχνική αυτή είναι πολύ σημαντική γιατί μπορεί να ανιχνεύσει μεταλλάξεις σε πολύ μικρό χρονικό διάστημα (λιγότερο από μια ημέρα), ενώ μέχρι σήμερα χρειάζονταν εβδομάδες ή και μήνες για να καταλήξουμε σε παρόμοια αποτελέσματα. Ωστόσο η τεχνική μέχρι σήμερα χρησιμοποιείται περισσότερο στην έρευνα παρά στην κλινική πρακτική εξαιτίας αφενός και της ανάγκης να διαδοθεί στην καθημερινή διαγνωστική των νοσοκομείων.

Τα δεδομένα που επιλέγονται συνήθως για περαιτέρω επεξεργασία είναι ο λόγος των λογαρίθμων της φωτεινότητας του κελιού (που όπως αναφέραμε αντιστοιχεί σε μια βάση) από την εικόνα που προέκυψε με χρήση πράσινου λέϊζερ, προς την φωτεινότητα του κελιού στην εικόνα που προέκυψε με χρήση κόκκινου λέϊζερ [1].

1.3 Επεξεργασία δεδομένων από μικροσυστοιχίες DNA

Αφού ολοκληρωθεί η προεπεξεργασία που έχει να κάνει κυρίως με επεξεργασία της εικόνας, προκύπτει ένας πίνακας (Σχήμα 1.2), όπου οι γραμμές αντιστοιχούν στα γονίδια και οι στήλες στα πειράματα. Οι αριθμοί όπως αναφέραμε είναι ο λογάριθμος των λόγων της φωτεινότητας στις δύο εικόνες. Πριν προχωρήσουμε στην εξακρίβωση των εξαρτήσεων μεταξύ των γονιδίων πρέπει να αντιμετωπίσουμε δύο σημαντικά προβλήματα. Πρώτον, όπως φαίνεται και στο Σχήμα 1.2, κατά κανόνα κάποιες τιμές από τα δεδομένα μας λείπουν.



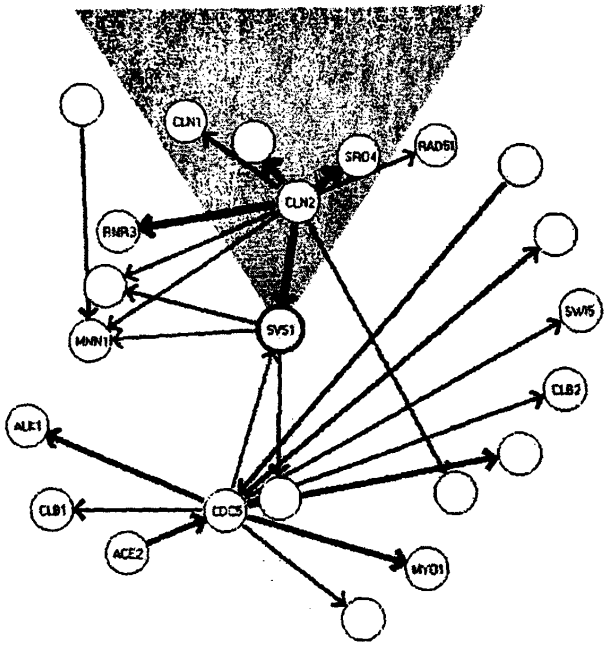
Σχήμα 1.2: Ένα μικρό μέρος από δεδομένα που προκύπτουν από μικροσυστοιχίες DNA

Επειδή η πλειοψηφία των μεθόδων που χρησιμοποιούνται για την εύρεση των ρυθμιστικών δικτύων, όπως τα Bayesian δίκτυα που θα δούμε στην συνέχεια, απαιτούν πλήρη δεδομένα, θα πρέπει να βρούμε κάποιο τρόπο να ανακτήσουμε αυτές τις τιμές ή έστω να υπολογίσουμε κάποιες προσεγγίσεις τους. Δεύτερον, ο αριθμός των γονιδίων είναι τεράστιος, γεγονός που απαγορεύει την επεξεργασία όλων των γονιδίων. Θα πρέπει να βρεθεί κάποιος τρόπος να μειώσουμε τον αριθμό των δεδομένων για επεξεργασία.

Για το πρώτο πρόβλημα η λύση δίνεται από αλγόριθμους και μεθοδολογίες ανάκτησης χαμένων τιμών, που θα εξετάσουμε στην συνέχεια.

Για το δεύτερο πρόβλημα οι λύσεις είναι δύο. Η πρώτη είναι να επιλέξουμε κάποια γονίδια ενδιαφέροντος, συνήθως χρησιμοποιώντας στατιστικά τεστ ώστε να διαπιστώσουμε αν παίζουν κάποιο ρυθμιστικό ρόλο και αξίζει να τα μελετήσουμε. Η δεύτερη προσέγγιση είναι η ομαδοποίηση των γονιδίων σε ομάδες που έχουν παρόμοια σχήματα έκφρασης και άρα υπόθετουμε έχουν παρόμοιο ρόλο στο συνολικό ρυθμιστικό δίκτυο. Στην συνέχεια αντί να μελετήσουμε τις αλληλεπιδράσεις μεταξύ των γονιδίων μελετάμε αυτές των ομάδων.

Αφού έχουμε λύσει τα παραπάνω προβλήματα, μπορούμε να προχωρήσουμε στο βασικό ζητούμενο. Να βρούμε δηλαδή τις εξαρτήσεις μεταξύ των γονιδίων και να κατασκευάσουμε το ρυθμιστικό δίκτυο. Στο συγκεκριμένο θέμα, η βιβλιογραφία είναι αρκετά εκτενής. Στο επόμενο κεφάλαιο παρουσιάζουμε τις βασικότερες μεθόδους. Αυτή η εργασία επικεντρώνεται σε μία από αυτές τις μεθόδους που βασίζεται στα Bayesian δίκτυα. Τα Bayesian δίκτυα θα τα εξετάσουμε αναλυτικά στην συνέχεια, αλλά μπορούμε προς το παρόν να τα περιγράψουμε σαν κατευθυνόμενα ακυκλικά γραφήματα, που κάθε κόμβος στο γράφημα αποτελεί μια μεταβλητή του προβλήματος, στο συγκεκριμένο πρόβλημα ένα γονίδιο, και κάθε ακμή αναπαριστά την εξάρτηση δύο μεταβλητών, π.χ. την εξάρτηση δύο γονιδίων (Σχήμα 1.3). Είναι σαφές και μόνο από αυτήν την απλοϊκή περιγραφή πόσο ταιριάζει το μοντέλο με το



Σχήμα 1.3: Ένα Bayesian δίκτυο για το γονίδιο SVS1 [2].

πρόβλημα που θέλουμε να λύσουμε.

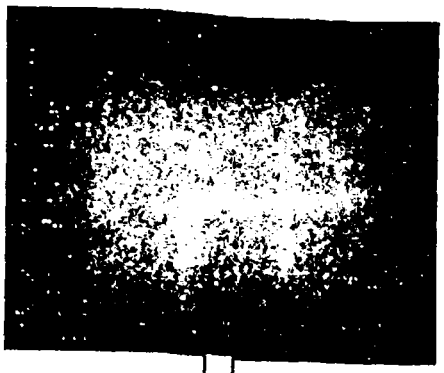
Σε αυτό το σημείο μπορούμε να συνοψίσουμε τα βασικά στάδια, που θα ακολουθήσουμε ως την εύρεση του ρυθμιστικού δικτύου (Bayesian δίκτυο) όπως αυτά φαίνονται στο Σχήμα 1.4.

1. Επεξεργασία εικόνας
2. Εξαγωγή δεδομένων
3. Προεπεξεργασία δεδομένων, ανάκτηση χαμένων τιμών
4. Ομαδοποίηση ή επιλογή συγκεκριμένων γονιδίων
5. Εύρεση ρυθμιστικού δικτύου

Τα δεδομένα για τα πειράματα όσο αναφορά την γονιδιακή έκφραση που παραθέτουμε στο τέλος προήρθαν από τρεις βάσεις ([3], [4], [5]) και συγκεντρωτικά έχουμε 460 διαφορετικά πειράματα για τα γονίδια της ζύμης.

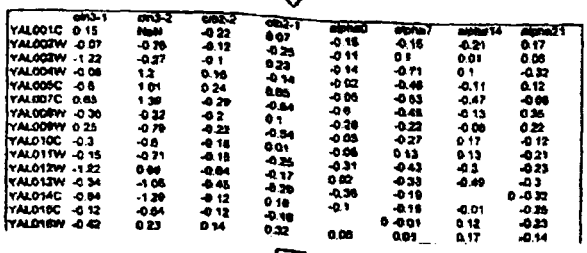
1.4 Στόχος

Ο βασικός στόχος της εργασίας αυτής είναι η εύρεση των εξαρτήσεων μεταξύ των γονιδίων με χρήση των δεδομένων των μικροσυστοιχιών DNA. Το βασικό εργαλείο που θα χρησιμοποιήσουμε είναι τα Bayesian δίκτυα. Ένα μεγάλο πρόβλημα των εφαρμογών για την εύρεση των ρυθμιστικών δικτύων στην βιβλιογραφία, είναι η αδυναμία λόγω υπολογιστικού κόστους να επεξεργαστούμε μεγάλα δίκτυα. Θα προσπαθήσουμε να ξεπεράσουμε αυτό



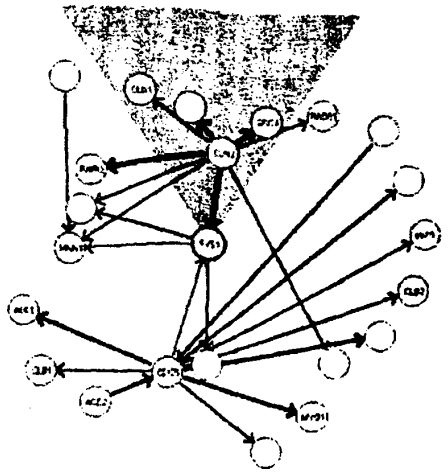
DNA CHIP

- ΕΠΕΞΕΡΓΑΣΙΑ ΕΙΚΟΝΑΣ
- ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ



ΔΕΔΟΜΕΝΑ

- ΑΝΑΚΤΗΣΗ ΧΑΜΕΝΩΝ ΤΙΜΩΝ
- ΟΜΑΔΟΠΟΙΗΣΗ Η' ΕΠΙΛΟΓΗ ΓΟΝΙΔΙΩΝ
- ΕΥΡΕΣΗ ΡΥΘΜΙΣΤΙΚΟΥ ΔΙΚΤΥΟΥ



ΡΥΘΜΙΣΤΙΚΟ ΔΙΚΤΥΟ ΓΟΝΙΔΙΩΝ

Σχήμα 1.4: Τα βήματα που ακολουθούνται για την κατασκευή του ρυθμιστικού δικτύου

το πρόβλημα με χρήση ενός παράλληλου συστήματος. Στην συνέχεια θα προσπαθήσουμε να εξακριβώσουμε αν ο Structural EM του N. Friedman [6] μπορεί να χρησιμοποιηθεί σε μεγάλα δίκτυα με παράλληλη υλοποίηση. Το πλεονέκτημα αυτής της μεθόδου είναι ότι χειρίζεται τις χαμένες τιμές στην διαδικασία της μάθησης οπότε δεν απαιτείται το στάδιο της ανάκτησης των χαμένων τιμών. Θα την συγκρίνουμε με την κλασική προσέγγιση όπου οι χαμένες τιμές ανακτώνται με κλασικές μεθόδους. Τέλος θα χρησιμοποιήσουμε και τις δύο προσεγγίσεις για να βρούμε εξαρτήσεις μεταξύ γονιδίων.

1.5 Δομή της Εργασίας

Στο κεφάλαιο 2 θα αναφερθούμε στην σχετική βιβλιογραφία. Στο κεφάλαιο 3 περιγράφουμε βασικές μεθόδους που χρησιμοποιούνται στην βιβλιογραφία για ανάκτηση χαμένων τιμών και ομαδοποίηση. Στο κεφάλαιο 4, κάνουμε μια εισαγωγή στα Bayesian δίκτυα και αναλύουμε τον τρόπο εκμάθησης παραμέτρων και δομής. Στο κεφάλαιο 5 επικεντρωνόμαστε στην περίπτωση που δεν έχουμε πλήρη δεδομένα και πως μπορούμε να αντιμετωπίσουμε αυτό το πρόβλημα στην περίπτωση των Bayesian δικτύων. Στην συνέχεια στο κεφάλαιο 6 περιγράφουμε την διαδικασία συμπερασματολογίας (inference) στα Bayesian δίκτυα, στο κεφάλαιο 7 περιγράφεται η παράλληλη υλοποίηση και τέλος στο κεφάλαιο 8 παραθέτουμε τα πειράματα μας.

Κεφαλαιο 2

Σχετικη Βιβλιογραφια

Η τεχνολογία των μικροσυστοιχιών, μπορεί να μας δώσει τιμές για την έκφραση ενός μεγάλου αριθμού γονιδίων. Λόγω αυτής της δυνατότητας, οι απαιτήσεις για αποτελέσματα στην εύρεση των εξαρτήσεων μεταξύ των γονιδίων έχουν αυξηθεί. Ο στόχος των ερευνητών είναι να εξακριβώσουν τις συνδέσεις του γενετικού δικτύου: για κάθε γονίδιο, θέλουμε να γνωρίζουμε ποια άλλα γονίδια αυτό επηρεάζει και με ποιο τρόπο. Σε αυτό το κεφάλαιο αναφερόμαστε στις βασικότερες εργασίες σχετικά με την ανακατασκεύη του γενετικού δικτύου, από τον σχετικά μικρό αριθμό δεδομένων που υπάρχουν διαθέσιμα.

2.1 Ομαδοποίηση

Στην περίπτωση που η έρευνα αφορά ένα μεγάλο αριθμό γονιδίων, η προσέγγιση που ακολουθείται από τους ερευνητές, είναι η εύρεση ομάδων γονιδίων με παρόμοια έκφραση [7]. Αυτό μπορεί να σημαίνει ότι η συμπεριφορά της έκφρασης σε διάφορες μεταλλάξεις ή ιστούς που έχουν μετρηθεί από μικροσυστοιχίες είναι παρόμοια [3]. Τα επιχειρήματα που μπορούμε να χρησιμοποιήσουμε, είναι ότι γονίδια εκφρασμένα με παρόμοιο τρόπο είναι πιθανόν να έχουν παρόμοιες λειτουργίες, ή απλά δεν υπάρχει καλύτερος τρόπος προς το παρόν για την ανάλυση των δεδομένων.

Παρόλο που θεωρητικά είναι δύσκολο από μια απλή ανάλυση συσχέτισης να καταλήξουμε σε δίκτυα αλληλεπίδρασης γονιδίων, πολλές εργασίες δείχνουν ότι η ομαδοποίηση των δεδομένων της έκφρασης των γονιδίων έχει σαν αποτέλεσμα, ομάδες γονιδίων να έχουν παρόμοια λειτουργία. Οι Eisen *et al.* [3] δημιούργησαν ένα δενδρόγραμμα από όλα τα γονίδια του Saccharomyces cerevisiae (ζύμη), η έκφραση του οποίου έχει μετρηθεί κάτω από

διαφορετικές συνθήκες. Από αυτά τα γονίδια, 35% έχουν μελετηθεί σε κάποια λεπτομέρεια και οι ομάδες που δίνει η ανάλυση τους, δείχνουν μεγάλα σύνολα από συσχετιζόμενα γονίδια. Συγκεκριμένα η ιεραρχική ομαδοποίηση και τα δεδρογράμματα περιγράφονται στην παράγραφο 4.3.4.

Οι Toronen *et al.* [8] χρησιμοποιούν self-orginizing map (SOM, Kohonen's map) για ομαδοποίηση των δεδομένων έκφρασης. Σε αυτόν το αλγόριθμο, επιλέγεται κάποια γεωμετρία για τους κόμβους, οι οποίοι τυχαία αντιστοιχίζονται σε ένα n-διάστατο χώρο. Στην συνέχεια εφαρμόζουν μια επαναληπτική διαδικασία που προσαρμόζει τις συντεταγμένες όλων των κόμβων έτσι ώστε να κινηθούν προς ένα τυχαία επιλεγμένο σημείο και όσο πιο κοντά βρίσκονται, τόσο πιο γρήγορα κινείται ο κόμβος. Τελικά, όλοι οι κόμβοι κατανέμονται στις ομάδες και δημιουργείται ένας χάρτης από το αρχικό πλέγμα όπου οι κόμβοι ήταν εναποθετημένοι. Από την στιγμή που οι Toronen *et al.* είχαν στην διάθεση τους μέρος από τα δεδομένα των Eisen *et al.* [3], τα δίκτυα που προκύπτουν από την ανάλυση τους, δεν περιέχουν τόσες ομάδες όσα τα συσχετισμένα ως προς την λειτουργία τους γονίδια και τα δενδρογράμματα της προηγούμενης παραγράφου. Για την ακρίβεια οι Toronen *et al.* [8] αναγνωρίζουν το πρόβλημα των εκφράσεων που δεν επηρεάζονται από τα συγκεκριμένα πειράματα. Αν υπήρχαν δεδομένα από ένα μόνο είδος πειραμάτων, τα γονίδια που δεν επηρεάζονται από αυτά τα πειράματα θα σχημάτιζαν μία ομάδα στο κέντρο του SOM. Αυτό σημαίνει ότι ο αλγόριθμος έχει πρόβλημα με την αναγνώριση σχημάτων έκφρασης που είναι αρκετά παρόμοια μεταξύ τους. Οι Toronen *et al.* που χρησιμοποίησαν επίσης τα SOM's για να σχηματίσουν ομάδες από δεδομένα μικροσυστοιχιών, υλοποίησαν ένα φίλτρο για γονίδια που δεν έχουν σημαντική αλλαγή στην έκφραση. Έτσι ενίσχυσαν την ευαισθησία του συστήματος και απέτρεψαν κόμβους να επηρεαστούν από μεγάλα σύνολα γονιδίων που δεν έχουν διακύμανση στην έκφραση. Το συμπέρασμα που προκύπτει από την ανάλυση των Toronen *et al.*, είναι ότι τα SOM αποτελούν ένα αξιόπιστο εργαλείο για την ομαδοποίηση συσχετισμένων γονιδίων.

Οι Hartuv *et al.* [9] ανέπτυξαν ένα καινούργιο αλγόριθμο ομαδοποίησης, βασισμένο σε μια γραφοθεωρητική προσέγγιση. Το κίνητρο για αυτήν την μελέτη ήταν η ανάγκη για ομαδοποίηση μια συλλογής από cDNA's με βάση τα αποτυπώματα ολιγονουκλεοτιδίων. Ωστόσο τα αποτελέσματα από αυτήν την εργασία είναι συγκρίσιμα με τα δεδομένα που λαμβάνονται από τις μικροσυστοιχίες. Ορίζεται ένα γράφημα ομοιότητας, στο οποίο οι ακμές υφίστανται αν η ομοιότητα των κόμβων που ενώνει η ακμή ξεπερνά κάποιο κατώφλι. Από αυτό το γράφημα, παίρνουμε μια «τομή» απομακρύνοντας ένα ελάχιστο αριθμό ακμών και υπολογίζεται η συνδεσιμότητα της τομής. Αν το υπογράφημα περιέχει περισσότερες από $n/2$ ακμές, ονομάζεται υψηλά συνδεδεμένο. Αυτά τα βήματα επαναλαμβάνονται, μέχρι τελικά, το σύνολο δεδομένων να ομαδοποιηθεί σε ένα αριθμό υψηλά συνδεδεμένων υπογραφημάτων, τις ομάδες. Οι συγγραφείς έδειξαν ότι ο αλγόριθμος έχει πολυωνυμική πολυπλοκότητα χαμηλού βαθμού, με μεγαλύτερη επιβάρυνση τον αλγόριθμο που εντοπίζει την ελάχιστη τομή. Τα αποτελέσματα του αλγορίθμου είναι καλά, τόσο για τα προσομειωμένα δεδομένα, όσο και για τα πραγματικά.

Η παράθεση των παραπάνω μελετών, δείχνει ότι η ομαδοποίηση των γονιδίων σύμφωνα με τα σχήματα έκφρασής τους μπορεί να περιέχει πληροφορία για την λειτουργία τους. Ωστόσο, όλοι οι ερευνητές τονίζουν την επιφυλακτικότητα τους, σχετικά με την διαδικασία ομαδοποίησης.

2.2 Εύρεση ρυθμιστικών δικτύων

Η πρώτη κατηγορία μεθόδων βασίζεται σε πίνακες βαρών. Ένας πίνακας βαρών αποτελείται από $n \times n$ τιμές βαρών, όπου η κάθε τιμή δείχνει την επιρροή ενός συγκεκριμένου γονιδίου σε κάποιο άλλο. Τα πλεονεκτήματα της μοντελοποίησης ρυθμιστικών δικτύων με πίνακες βαρών αναφέρονται από τους Weaner *et al.* [10], οι οποίοι παρουσίασαν ένα αλγόριθμο (TReMM: Transcription REgulation Modelled with Matrices) για ένα κλασσικό παράδειγμα ενός μοντέλου με πίνακες βαρών. Τα βάρη W_{ij} αναπαριστούν τη επιρροή του γονιδίου i στο γονίδιο j, και η πλήρης είσοδος σε κάποιο γονίδιο j δίνεται από το άθροισμα όλων των εισόδων των γονιδίων i, πολλαπλασιαζόμενα με το βάρος τους. Το αποτέλεσμα αυτού του υπολογισμού, εισάγεται σε μια εξίσωση κανονικοποίησης η οποία δίνει σαν έξοδο μια τιμή έκφρασης από 0 ως 1 για κάθε γονίδιο j. Η εύρεση των τιμών γίνεται συνήθως με κλασσικούς μηχανισμούς μάθησης, όπως simulated annealing [11] και γενετικούς αλγοριθμούς. Όταν αρχίζουμε μια μελέτη ενός ρυθμιστικού δικτύου σε ένα συγκεκριμένο βιολογικό σύστημα, τα $n \times n$ βάρη που συνθέτουν τον πίνακα είναι άγνωστα. Η ιδέα είναι, ότι τα βάρη μπορούν να προσεγγιστούν από τα δεδομένα έκφρασης στην διαδικασία της κατασκευής του δικτύου.

Είναι σαφές, πως ένα μοντέλο με πίνακα βαρών λαμβάνει υπόψη τις αλληλεπιδράσεις μεταξύ όλων των συνδυασμών των γονιδίων, πολλές από τις οποίες έχουν τιμή 0. Επειδή δεν είναι εκ των προτέρων γνωστό ποιες είναι 0, το υπολογιστικό κόστος είναι μεγάλο.

Για να αυξήσουν τον αριθμό των δεδομένων από τα οποία ξεκινάμε την ανακατασκευή του γενετικού δικτύου, οι D'Haeseleer *et al.* [12] άρχισαν την ανάλυση τους με τον υπολογισμό μιας μη γραμμικής καμπύλης παρεμβολής στις χρονοακολουθίες έκφρασης των γονιδίων. Έτσι χρησιμοποίησαν ακολουθίες δεδομένων από τρία διαφορετικά πειράματα, με πολύ διαφορετικές κλίμακες χρόνου (τα χρονικά διαστήματα μεταξύ των πειραμάτων ποικίλουν από μισή ώρα μέχρι δύο μήνες). Κατάφεραν να συνδυάσουν δεδομένα από τις ακολουθίες που είχαν παρεμβάλει με κυβική παρεμβολή στον λογάριθμο του επιπέδου έκφρασης. Το επόμενο βήμα είναι η εύρεση των τιμών στον πίνακα βαρών όπως περιγράψαμε στην προηγούμενη παράγραφο. Αυτά τα βάρη αλληλεπίδρασης προσεγγίζονται με χρήση ελάχιστων τετραγώνων στις ακολουθίες που έχουν παρεμβάλει. Το αποτέλεσμα αυτής της προσέγγισης είναι ένα ακριβές ταίριασμα στα δεδομένα. Ταυτόχρονα, παρέχουν και κάποια σχόλια για την μέθοδο. Πρώτον, ο μηχανισμός αυτός δεν ελαχιστοποιεί τον αριθμό των αλληλεπιδράσεων μεταξύ των γονιδίων: κάθε γονίδιο μοντελοποιείται από τον σταθμισμένο μέσο όρο όλων των άλλων γονιδίων. Δεύτερον, το απλό γραμμικό προσθετικό μοντέλο της αλληλεπίδρασης των γονιδίων, μπορεί να αφομοιώσει τις κύριες σχέσεις του συστήματος.

Τέλος επειδή τα δεδομένα που είχαν στην διάθεση τους οι ερευνητές δεν ήταν όμοια κατανεμημένα στον χρόνο, μεγαλύτερο βάρος δινόταν στα πιο αραιά κατανεμημένα σημεία στα δεδομένα.

Οι Chen και Church [13] πρότειναν ένα μοντέλο διαφορικών εξισώσεων για την έκφραση των γονιδίων και ανέπτυξαν δύο μεθόδους για την κατασκευή του δικτύου από τα δεδομένα. Έκαναν ένα αριθμό υποθέσεων, μεταξύ των οποίων, ότι υπάρχει μια γραμμική συνάρτηση μεταγραφής για κάθε γονίδιο και ότι η οπισθοτροφοδότηση της μετάφρασης κάθε γονιδίου έχει σαν αποτέλεσμα τους ρυθμούς μεταγραφής. Επίσης αγνοούνται άλλες οπισθο-τροφοδοτήσεις όπως του mRNA στα γονίδια, αφού συνυπολογίζονται σε αυτές των πρωτεϊνών. Οι συγγραφείς υποθέτουν ότι ο όλος μηχανισμός μεταγραφής είναι σχετικά ευσταθής (για ένα τουλάχιστον μικρό χρονικό διάστημα), ώστε η οπισθο-τροφοδότηση από τις πρωτείνες στα mRNA να μην έχει καμία επίδραση. Κάθε mRNA και μόριο πρωτείνης μειώνεται τυχαία και τα στοιχεία του ανακυκλώνονται στο κύτταρο. Διακρίνουν στο μοντέλο τους, την μεταγραφή, την μετάφραση και την μείωση του RNA και των πρωτεινών. Για την εύρεση των παραμέτρων χρησιμοποιούν μετασχηματισμό Fourier, με την υπόθεση ότι το σύστημα είναι ευσταθές. Με την υπόθεση ότι το μοντέλο μεταγραφής είναι γραμμικό και το σύστημα πρέπει να είναι ευσταθές, οι Chen *et al.* κατάφεραν να μειώσουν δραστικά την διάσταση του χώρου των παραμέτρων. Χάρη σε αυτήν την μείωση, μπορούν και εισάγουν περισσότερα χαρακτηριστικά στο μοντέλο τους, σε σχέση με πιο απλά μοντέλα. Ωστόσο, οι συγγραφείς δεν αναφέρουν πόση επιρροή έχει στα αποτελέσματα το πολυπλοκότερο αυτό μοντέλο, δηλαδή το πόσο σημαντικές είναι οι επιπλέον υποθέσεις.

Οι Shmulevich *et al.* [14] χρησιμοποίησαν Boolean δίκτυα για την εύρεση των ρυθμιστικών δικτύων. Τα Boolean δίκτυα, που εισήχθησαν από τον Kaufmann [15] και μελετήθηκαν εκτενέστερα στα [16, 17], προσφέρουν μια ελκυστική διακριτού χρόνου, boolean αναπαράσταση του μοντέλου για την έκφραση των γονιδίων. Σε αυτά τα μοντέλα διακριτοποιούμε την έκφραση των γονιδίων στις καταστάσεις ON (εκφρασμένα) και OFF (μη εκφρασμένα). Η έκφραση ενός γονιδίου την χρονική στιγμή $t+1$ μοντελοποιείται από μια Boolean συνάρτηση της οποίας οι είσοδοι είναι το πολύ k γονίδια την στιγμή t. Τυπικά $k \ll n$, όπου n ο αριθμός των γονιδίων. Ονομάζουμε αυτή την κλάση των μοντέλων, Boolean δίκτυα ($BN(n,k)$). Μερικά από τα ελκυστικά χαρακτηριστικά αυτών των δικτύων είναι η απλότητα της λογικής τους και η διαφάνεια, όπως και η ποικιλία αλγορίθμων για την αυτόματη εύρεση του μοντέλου από τα δεδομένα. Πιο λεπτομερή περιγραφή υπάρχει στην αναφορά [18]. Στην εργασία [19] τα Boolean δίκτυα επεκτάθηκαν ώστε να μπορούν να χειριστούν και σχέσεις που έχουν διάρκεια μεγαλύτερη της μίας χρονικής περιόδου, και ονομάστηκαν *Temporal Boolean Networks (TBN).*

Οι Friedman *et al.* [2] χρησιμοποίησαν Bayesian δίκτυα για την μοντελοποίηση του γενετικού δικτύου. Τα Bayesian δίκτυα είναι ένα πολύ χρήσιμο στατιστικό εργαλείο, στο οποίο τα τελευταία δέκα χρόνια έχει γίνει μεγάλη έρευνα και οι εφαρμογές του συνεχώς αυξάνονται. Θα τα περιγράψουμε αναλυτικά στο κεφάλαιο 4. Όπως ήδη αναφέραμε ένα

από τα σημαντικότερα προβλήματα στην ανακατασκεύη των γενετικών δικτύων είναι ο τεράστιος αριθμός γονιδίων που αντιστοιχούν σε κόμβους στο γενετικό δίκτυο που θέλουμε να κατασκευάσουμε. Για παράδειγμα αν έχουμε n γονίδια τα πιθανά γενετικά δίκτυα είναι περίπου $n!$ αν θεωρήσουμε ότι κάθε γονίδιο μπορεί να συσχετίζεται με όλα τα άλλα. Οι Friedman *et al.* για να αντιμετωπίσουν αυτό το πρόβλημα πρότειναν έναν αλγόριθμο τον οποίο ονόμασαν Sparse Candidate. Στην ουσία για κάθε κόμβο-γονίδιο του δικτύου βρίσκουν ένα μικρό υποσύνολο από το τεράστιο σύνολο των γονιδίων που είναι πιο πιθανό να ανήκουν στο σύνολο των γονιδίων που το επηρεάζουν. Για την μοντελοποίηση της έκφρασης των γονιδίων χρησιμοποίησαν τόσο διακριτές όσο και συνεχείς κατανομές. Εφαρμόζουν τον αλγόριθμο τους στα δεδομένα του Spellman [20] για το γονιδίωμα της ζύμης (*S. cerevisiae*) στον κύκλο του κυττάρου. Επίλεξαν 800 γονίδια που η έκφραση τους είχε μεγάλη διακύμανση στα δεδομένα. Τα αποτελέσματα τους έδειξαν ότι μπορούμε να πάρουμε πολύπλοκες δομές ακόμα και με λίγα δεδομένα ($\sim$ 73). Οι Spirtes *et al.* [21] ανέλυσαν τα προβλήματα που παρουσιάζονται όταν χρησιμοποιούμε Bayesian δίκτυα για να βρούμε τις εξαρτήσεις μεταξύ των γονιδίων.

Οι Imoto *et al.* [22] χρησιμοποίησαν μη παραμετρικές στατιστικές μεθόδους παλινδρόμησης με βάση τα B-Splines και τα Bayesian δίκτυα για την μοντελοποίηση μη γραμμικών σχέσεων μεταξύ των γονιδίων. Με βάση αυτό το μοντέλο ανέπτυξαν ένα νέο κριτήριο επιλογής δικτύων και αξιολόγησαν την μέθοδό τους σε 100 γονίδια της ζύμης.

Οι Murphy και Mian [23] χρησιμοποίησαν δυναμικά Bayesian δίκτυα και έδειξαν ότι το γραμμικό μοντέλο των D'Haeseleer *et al.* [12], το μη γραμμικό μοντέλο του Weaver *et al.* [10] και τα Boolean δίκτυα είναι υποκατηγορίες των δυναμικών Bayesian δικτύων.

Οι Hartemink *et al.* [24] βασίστηκαν σε ένα μοντέλο που επιτρέπει τον χειρισμό κρυμμένων μεταβλητών και βασίζεται στα Bayesian δίκτυα και τις επεκτάσεις τους. Για να αξιολογήσουν την προσέγγιση τους, χρησιμοποίησαν δεδομένα για τα γονίδια του ρυθμιστικού δικτύου του μεταβολισμού της γαλακτόζης στη ζύμη (*S. cerevisiae*). Επικεντρώθηκαν κυρίως στην αξιολόγηση των δικτύων που προκύπτουν.

Οι Ong *et al.* [25] περιγράφουν μια προσέγγιση που χειρίζεται δεδομένα από χρονοακολουθίες και μπορεί να χειριστεί κύκλους επανατροφοδότησης καθώς και περιβαλλοντολογικές συνθήκες σαν κρυμμένες μεταβλητές. Επίσης παρουσιάζουν ένα καινούργιο τρόπο για τον συνδυασμό της εκ των προτέρων γνώσης και των δεδομένων. Δοκίμασαν την μέθοδο τους σε δεδομένα έκφρασης που μετρήθηκαν σε φυσιολογικές αλλαγές που επηρεάζουν τον μεταβολισμό του tryptophan στο *E.coli*. Τα αποτελέσματα δείχνουν ότι η μέθοδος είναι ικανή να βρίσκει σχέσεις μεταξύ συνόλων γονιδίων που πραγματικά συσχετίζονται.

Οι Perrin *et al.* [26] χρησιμοποίησαν μία προσέγγιση με γραμμικά δυναμικά Bayesian δίκτυα συνεχών μεταβλητών. Ο αλγόριθμος εκπαίδευσης μεγιστοποιεί την πιθανοφάνεια με κάποιους περιορισμούς που ευνοούν «χρήσιμες» συνδέσεις. Αυτή η προσέγγιση υλοποιήθηκε

για ένα καινούργιο δυναμικό μοντέλο αλληλεπιδράσεων που το ονόμασαν inertial model. Τα αποτελέσματα τους στο S.O.S DNA repair δίκτυο του *E.coli* είναι αρκετά ενθαρρυντικά.

Στις παραπάνω εργασίες, αλλά και στις περισσότερες εργασίες που αναφέρονται στην εύρεση των ρυθμιστικών δικτύων ακολουθούνται γενικά τα ακόλουθα βήματα. Πρώτα γίνεται μία απλή προεπεξεργασία στα δεδομένα. Στην συνέχεια χρησιμοποιούνται αλγόριθμοι για την ανάκτηση των χαμένων τιμών και τέλος εφαρμόζονται μεθόδοι για την κατασκευή των ρυθμιστικών δικτύων. Όπως αναφέρθηκε στην σύνοψη της βιβλιογραφίας οι αλγόριθμοι ομαδοποίησης μπορούν να χρησιμοποιηθούν για την εύρεση των ομάδων γονιδίων που έχουν παρόμοια έκφραση ώστε να μελετηθεί η λειτουργία των γονιδίων σύμφωνα με κάποια γονίδια της ομάδας των οποίων η λειτουργία είναι παρόμοια. Την υπόθεση ότι σε μια ομάδα τα γονίδια έχουν παρόμοια λειτουργία μπορούμε να την εκμεταλλευτούμε για να μειώσουμε το κόστος στην εύρεση των ρυθμιστικών δικτύων. Αντί να κατασκευάσουμε το δίκτυο με τον τεράστιο αριθμό των γονιδίων κατασκευάζουμε ένα δίκτυο με πολύ μικρότερο αριθμό ομάδων.

Κεφαλαιο 3

Προεπεξεργασια

3.1 Προεπεξεργασία δεδομένων

Η πιο συνηθισμένη μορφή προεπεξεργασίας που εφαρμόζεται για την εύρεση των ρυθμιστικών δικτύων από τις μικροσυστοιχίες είναι μια απλή γραμμική κλιμάκωση των μεταβλητών [1]. Αυτό είναι συνήθως χρήσιμο όταν οι μεταβλητές του προβλήματος μας έχουν τιμές που διαφέρουν σημαντικά. Χρησιμοποιείται ένας γραμμικός μετασχηματισμός, ώστε όλες οι μεταβλητές να έχουν παρόμοιες τιμές. Για να το πετύχουμε αυτό, από κάθε μεταβλητή X_i αφαιρούμε το μέσο της και διαιρούμε με την διασπορά που ορίζονται ως:

$$\bar{X}_i = \frac{1}{N}\sum_{j=1}^{N} x_{ij} \tag{3.1}$$

$$\sigma_i^2 = \frac{1}{N-1}\sum_{j=1}^{N}(x_{ij} - \bar{X}_i)^2 \tag{3.2}$$

όπου $j = 1, ..., N$, τα δείγματα μας. Με x_{ij} δηλώνεται η j-οστή τιμή της X_i στα δεδομένα, με $\bar{X}_i$ η μέση τιμή και με σ_i η διασπορά της μεταβλητής X_i. Η j-οστή τιμή της κλιμακωμένης μεταβλητής $\tilde{X}_i$, ορίζεται ως εξής:

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{X}_i}{\sigma_i} \tag{3.3}$$

Οι κλιμακωμένες μεταβλητές έχουν μέσο μηδέν και διασπορά ίση με ένα.

3.2 Ανάκτηση Χαμένων τιμών

Το πρώτο πρόβλημα που παρουσιάζεται είναι ότι τα δεδομένα δεν είναι πλήρη, δηλαδή σε κάποιο πείραμα η έκφραση κάποιου γονιδίου μπορεί να λείπει (missing value). Αυτό οφείλεται τόσο στην μεθοδολογία και στα συστήματα συλλογής δεδομένων (μικροσυστοιχίες DNA), όσο και στην φύση των δεδομένων, αφού είναι πιθανόν κάποιο γονίδιο, υπό κάποιες συγκεκριμένες συνθήκες να μην εκφράζεται. Στο πρώτο βήμα, πρέπει να ανακτηθούν οι ελλιπείς (ή χαμένες) τιμές. Εδώ πρέπει να αναφέρουμε ότι ο χειρισμός των ελλιπών τιμών μπορεί να γίνει και σε επόμενο βήμα της επεξεργασίας, το οποίο αναλύουμε στο κεφάλαιο 5.

Η πρώτη και πιο απλοϊκή προσέγγιση που χρησιμοποιείται, είναι η ανάκτηση των χαμένων τιμών αντικαθιστώντας τις με μηδενικά ή με το μέσο της αντίστοιχης γραμμής [27]. Δύο πιο εξελιγμένες προσεγγίσεις είναι αυτές των K-NN και SVD [27].

Οι Oba *et al.* [28] χρησιμοποίησαν την Bayesian PCA μέθοδο για την ανάκτηση χαμένων τιμών, οι Kim *et al.* [29] χρησιμοποίησαν μέθοδο ελαχίστων τετραγώνων και EM. Και οι δύο μεθόδοι είχαν σχετικά καλύτερα αποτελέσματα από τον K-NN στις περισσότερες περιπτώσεις αλλά λόγω της απλότητας του, ο K-NN είναι η μέθοδος που χρησιμοποιείται συχνότερα. Παρακάτω παρουσιάζουμε πιο αναλυτικά τις μεθόδους με χρήση K-NN, SVD και παλινδρόμησης.

3.2.1 Ανάκτηση χαμένων τιμών με SVD

Ανάκτηση με SVD χρησιμοποιώντας πλήρη δεδομένα

Η πρώτη μέθοδος που εξετάζουμε και έχει εφαρμοστεί για την ανάκτηση χαμένων τιμών είναι η SVD [27]. Έστω X ο $n \times N$ πίνακας με τα δεδομένα, όπου n αριθμός των μεταβλητών (γονιδίων) και N ο αριθμός των πειραμάτων. Έστω τώρα X^c το υποσύνολο των μεταβλητών που δεν έχουν ελλιπείς τιμές και X^m οι υπόλοιπες που έχουν τουλάχιστον μία χαμένη τιμή. Θεωρούμε τον truncated SVD του X^c:

$$\hat{X}_J^c = U_J D_J V_J^T, \tag{3.4}$$

όπου D_J είναι ένας διαγώνιος πίνακας που περιέχει τις $J \leq N$ singular τιμές του X^c, V_J και U_J οι αντίστοιχοι ορθογώνιοι πίνακες των J δεξιών και αριστερών singular διανυσμάτων. Ο SVD τάξης J μπορεί να ερμηνευθεί από διαφορετικές οπτικές γωνίες.

Μία ερμηνεία που μας βολεύει για το πρόβλημα, είναι αυτή της παλινδρόμησης (regression). Έστω X_i κάποια γραμμή του X^c. Θεωρούμε την παλινδρόμηση ελαχίστων τετραγώνων των N τιμών της X_i στα singular διανύσματα $v_1, v_2, ..., v_j$, που το καθένα είναι διάνυσμα διάστασης N και β το διάνυσμα των συντελεστών παλινδρόμησης. Η παλινδρόμηση αυτή λύνει το παρακάτω πρόβλημα των ελαχίστων τετραγώνων

$$\min_{\beta} \| X_i - V_J \beta \|^2 = \min_{\beta} \sum_{l=1}^{N} (x_{il} - \sum_{j=1}^{J} v_{lj} \beta_j)^2, \tag{3.5}$$

όπου v_{lj} το στοιχείο l, j του πίνακα V_J. Η λύση είναι $\hat{\beta} = (V_J^T V_J)^{-1} V_J^T X_i = V_J^T X_i$ (ο V_J ορθογώνιος άρα $V_J^T V_J = I$) και μας δίνει τιμές $\hat{x} = V_J \hat{b}$. Έτσι σύμφωνα και με την σχέση (3.4) το $X^c V_J = U_J D_J$ δίνει όλους του συντελεστές παλινδρόμησης για όλες τις γραμμές και το $\hat{X}_c = U_J D_J V_J^T$ όλες τις τιμές. Αφού έχει βρεθεί ο V_J, ο SVD προσεγγίζει κάθε γραμμή του X^c με το διάνυσμα που δίνεται με παλινδρόμηση στον V_J.

Αυτό επίσης μας λέει ότι για κάθε γραμμή X_i του X^m με κάποια ελλιπή στοιχεία, μπορούν να ανακτηθούν τα χαμένα στοιχεία με μια παρόμοια παλινδρόμηση:

$$\min_{\beta} \sum_{l_{\text{μη-ελλιπές}}} (x_{il} - \sum_{j=1}^{J} v_{lj}\beta_j)^2. \tag{3.6}$$

Έστω V_j^* η μειωμένη εκδοχή του V_J με αφαιρεμένες τις κατάλληλες γραμμές δηλαδή αυτές που περιέχουν ελλιπείς τιμές. Η λύση στην εξίσωση (3.6) γίνεται $\hat{\beta} = (V_j^{*T} V_j^*)^{-1} V_j^{*T} X_i^*$, και οι εκτιμήσεις για τις ελλιπείς τιμές είναι $V_j^{(*)}\hat{\beta}$ όπου $V_j^{(*)}$ το συμπλήρωμα του V_j^* στον πίνακα V_J. Σημειώνουμε ότι οι στήλες V_j^* δεν είναι πλέον ορθογώνιες.

Ανάκτηση με SVD χρησιμοποιώντας όλα τα δεδομένα

Η προηγούμενη προσέγγιση μπορεί να εφαρμοστεί μόνο όταν τα πλήρη δεδομένα είναι αρκετά και τα ελλιπή δεν συνεισφέρουν στην βάση του SVD. Αυτό φυσικά δεν ισχύει πάντα. Η προσέγγιση που περιγράφεται στην συνέχεια χρησιμοποιεί όλα τα δεδομένα. Για να λυθεί το πρόβλημα της ανάκτησης των χαμένων τιμών, ανάγεται στην λύση του παρακάτω προβλήματος

$$\min_{U_J, V_J, D_J} \| X - m1^T - U_J D_J V_J^T \|^* . \tag{3.7}$$

Όπου $\| \cdot \|^*$, μία ειδική τετραγωνική νόρμα, η οποία αθροίζει το τετράγωνο όλων των στοιχείων αγνοώντας τα στοιχεία όπου ο X έχει ελλιπείς τιμές. Ο m είναι ένα διάνυσμα με το μέσο κάθε γραμμής του X και 1 ένα διάνυσμα με μονάδες διάστασης N. Αν δεν υπάρχουν ελλιπείς τιμές, η λύση είναι τετριμμένη: το m είναι το διάνυσμα των μέσων των γραμμών του X και οι U_j, V_j και D_j δίνονται από τον J τάξης SVD του κεντρικοποιημένου X. Αφού η λύση τάξης J βρεθεί για το πρόβλημα, χρησιμοποιείται για να «γεμίσουν» οι ελλιπείς τιμές του X. Η διαδικασία που ακολουθείται περιγράφεται από τον Αλγόριθμο 1.

Αλγόριθμος 1 Αλγόριθμος ανάκτησης χαμένων τιμών με SVD

$i := 0$

Θέτονται οι ελλιπείς τιμές ίσες με το μέσο των μη χαμένων τιμών κάθε γραμμής, παράγοντας έναν πλήρη πίνακα $X^{(0)}$.

repeat

Υπολογίζεται η SVD λύση για το πρόβλημα (3.7), για τον πλήρη πίνακα και παράγεται ο X^{i+1} αντικαθιστώντας τις ελλιπείς τιμές στον X με αυτές που μας δίνει η λύση.

$i := i + 1$

until Ο όρος $\| X^{(i)} - X^{(i+1)} \| / \| X^{(i)} \|$ να γίνει μικρότερος από κάποιο όριο.

3.2.2 K-NN

Αυτή είναι η πλέον συχνή προσέγγιση στο πρόβλημα που δίνει πολύ καλά αποτελέσματα [27]. Η περιγραφή του αλγορίθμου συνοψίζεται στα εξής βήματα:

1. Υπολογίζεται η Ευκλείδια απόσταση μεταξύ του γονιδίου X_i και των άλλων γονιδίων στο X^c, χρησιμοποιώντας μόνο τα σημεία που δεν είναι ελλιπή στο X_i. Αναγνωρίζονται τα K κοντινότερα.

2. Υπολογίζονται οι ελλιπείς τιμές του X_i παίρνοντας τον μέσο όρο των αντίστοιχων σημείων στα K κοντινότερα.

3.2.3 Ανάκτηση με παλινδρόμηση

Αυτή η τεχνική είναι μια κλασική *EM* προσέγγιση για την εύρεση μέσων και πινάκων συμμεταβλητότητας για πολυδιάστατες Gaussian κατανομές, για ελλιπή δεδομένα [29]. Οι προβλεπόμενες τιμές για χαμένα στοιχεία παράγονται στην διαδικασία του αλγορίθμου. Η βασική ιδέα είναι, για κάθε j, να χρησιμοποιηθεί παλινδρόμηση της στήλης του j, σε κάθε άλλη στήλη εκτός αυτής του j, για να βρούμε τις ελλιπείς τιμές στην j. Πιο συγκεκριμένα, για κάθε στήλη j:

1. Αφαιρούνται οι γραμμές του X οι οποίοι περιέχουν ελλιπείς τιμές στην στήλη j.

2. Πραγματοποιείται παλινδρόμηση της «καθαρής» πλέον στήλης σε κάθε άλλη γραμμή του (μειωμένου) X.

3. Χρησιμοποιούνται οι συντελεστές της παλινδρόμησης για να γίνουν προβλέψεις σχετικά με τις ελλιπείς τιμές της στήλης j.

Επειδή θα υπάρχουν ελλιπείς τιμές και στις άλλες στήλες έχει δημιουργηθεί μία παραλλαγή της παραπάνω διαδικασίας. Χρησιμοποιείται μια επαναληπτική διαδικασία (EM), όπου έχουν αντικατασταθεί οι ελλιπείς τιμές με υποθέσεις (αρχικά το μέσο της κάθε γραμμής) και αυτές οι υποθέσεις ανανεώνονται καθώς η διαδικασία προχωράει.

Γενικά η τελευταία προσέγγιση έχει καλύτερη απόδοση σε σχέση με τις K-NN και SVD. Παρόλα αυτά είναι πολύ πιο αργή. Μεταξύ K-NN και SVD σε μελέτες που έχουν γίνει [27], έχει δειχθεί πειραματικά ότι η μέθοδος K-NN έχει ελαφρώς καλύτερη απόδοση αφού η SVD έχει το μειονέκτημα ότι οι singular τιμές κυριαρχούνται από τα πιο κοινά σχήματα έκφρασης, οπότε αν έχουμε ελλιπείς τιμές σε σπάνιας μορφής σχήματα έκφρασης δεν μπορούν να προβλεφθούν.

3.3 Ομαδοποίηση

Ένας σημαντικός στόχος της ανάλυσης του μεγάλου όγκου δεδομένων από τις μικροσυστοιχίες DNA, είναι να ξεχωρίσουν βασικές μορφές τις γονιδιακής έκφρασης, οι οποίες

δίνουν την κύρια πληροφορία για την βιολογία των δειγμάτων. Γονίδια με παρόμοια έκφραση σε διάφορες συνθήκες ρυθμίζονται συνήθως από τους ίδιους παράγοντες. Σαν ένα εργαλείο περιγραφής, η ομαδοποίηση των σχημάτων έκφρασης μπορεί να μας αποκαλύψει τέτοιες σχέσεις. Τα ποσοτικά επίπεδα έκφρασης n γονιδίων σε d συνθήκες μπορούν να θεωρηθούν σαν n σημεία σε ένα d-διάστατο χώρο. Οι αλγόριθμοι ομαδοποίησης βάζουν σε ομάδες σημεία που είναι κοντά στον d-διάστασης χώρο. Η ομαδοποίηση έχει αποδειχθεί πολύ χρήσιμη σε μελέτες του καρκίνου [30].

3.3.1 Μέτρα απόστασης και ομοιότητας

Η «ομοιότητα» μεταξύ γονιδίων γίνεται συγκεκριμένη αν έχουμε ορίσει ένα μέτρο απόστασης ή ένα μέτρο ομοιότητας που μπορεί να περιγράψει ποσοτικά το πόσο όμοιες ή ανόμοιες είναι δύο εκφράσεις γονιδίων μεταξύ τους. Για n γονίδια, για κάθε ζευγάρι (X_k, X_l) από τα $(\frac{n(n-1)}{2})$ ζευγάρια γονιδίων μπορεί να μετρηθεί η ομοιότητα του για τα επίπεδα έκφρασης υπό τις N συνθήκες. Ένα ευρέως χρησιμοποιούμενο μέτρο είναι η ευκλείδια απόσταση:

$$d_2(X_k, X_l) = \sqrt{\sum_{j=1}^{N}(X_{kj} - X_{lj})^2}. \tag{3.8}$$

Ένα άλλο βολικό μέτρο είναι ο συντελεστής συσχέτισης που υπολογίζει πόσο συσχετισμένα είναι τα επίπεδα έκφρασης των γονιδίων X_k και X_l υπό N διαφορετικές συνθήκες:

$$R(X_k, X_l) = \frac{\sum_{j=1}^{N}(X_{kj} - \overline{X}_k)(X_{lj} - \overline{X}_l)}{\sqrt{\sum_{i=1}^{N}(X_{kj} - \overline{X}_k)^2}\sqrt{\sum_{i=1}^{N}(X_{lj} - \overline{X}_l)^2}}. \tag{3.9}$$

3.3.2 K-means

Μια ευρέως χρησιμοποιούμενη μέθοδος ομαδοποίησης είναι ο αλγόριθμος K-means και διάφορες παραλλαγές του [31, 32]. Έχει το πλεονέκτημα ότι δεν επιβάλει αυστηρή σχέση σε κάθε γονίδιο, το οποίο μπορεί να είναι προβληματικό από την στιγμή που δεν υπάρχει απόλυτα διαδοχική σχέση μεταξύ των σχημάτων έκφρασης. Σε αυτήν την μέθοδο, τα γονίδια ταξινομούνται σαν να ανήκουν σε μία από τις k ομάδες. Η συμμετοχή κάθε γονιδίου σε μία από τις ομάδες $C_1, C_2, \ldots, C_k$ με κέντρο $a_1, a_2, \ldots, a_k \in \mathbb{R}^N$, υπολογίζεται αναθέτοντας κάθε γονίδιο i σύμφωνα με το σχήμα έκφρασης του X_i, στην ομάδα με το κοντινότερο κέντρο [3]. Ο στόχος είναι η εύρεση εμπειρικά, των βέλτιστων κέντρων $a_1, ..., a_k$, ώστε το σφάλμα

$$E = \frac{1}{n}\sum_{i=1}^{n}\min_{0\le j\le k} \| X_i - a_j \|^2, \tag{3.10}$$

να ελαχιστοποιείται, όπου n ο αριθμός των μεταβλητών. Αυτό επιτυγχάνεται με την ακόλουθη διαδικασία:

Αλγόριθμος 2 K-means Αλγόριθμος

$i := 0$

Αναθέτονται τιμές στα αρχικά κέντρα $a_1^{(0)}, ..., a_k^{(0)}$ με αυθαίρετο τρόπο.

repeat

1. Ομαδοποιούνται οι μεταβλητές(γονίδια) $X_1, ..., X_n$, στις k ομάδες.

Για X_j, $j \in [1, ..., n]$

Αν $\| X_j - a_m \|^2 \leq \| X_j - a_l \|^2$ για κάθε $l \neq m$,

αναθέτεται το X_j στην m ομάδα.

2. Ανανεώνονται τα κέντρα των ομάδων

$$a_m^{(i+1)} = \sum_{j:X_j \in C_m^{(i)}} X_j / |C_m^{(i)}|$$

$i := i + 1$

until Να μη γίνουν σημαντικές αλλαγές στα κέντρα των ομάδων (Σύμφωνα με κάποιο κριτήριο).

3.3.3 Ιεραρχική Ομαδοποίηση

Αυτή η μέθοδος ομαδοποιεί γονίδια σε ένα δέντρο ή δενδρόγραμμα [3,33]. Στην αρχή, κάθε γονίδιο αποτελεί μια ξεχωριστή ομάδα. Ξεκινώντας από τις n ομάδες, οι δύο ομάδες με την μικρότερη απόσταση συγχωνεύονται σε μια νέα ομάδα και οι ομάδες ανανεώνονται. Αυτή η διαδικασία επαναλαμβάνεται μέχρι να προκύψει μια ομάδα.

Για να ανανεωθεί ο πίνακας αποστάσεων όταν δύο ομάδες C_i και C_j συγχωνεύονται σε μία νέα C_q, η ερώτηση κλειδί είναι πώς να οριστεί η απόσταση της νέας ομάδας C_q και των άλλων ομάδων.

1. Στην *single linkage* προσέγγιση, η απόσταση της ομάδας C_q από μια άλλη υπάρχουσα ομάδα C_s, υπολογίζεται ως εξής:

$$d(C_q, C_s) = min(d(C_i, C_s), d(C_j, C_s))$$

όπου d το μέτρο απόστασης ή ομοιότητας που χρησιμοποιούμε.

2. Στην *complete linkage* προσέγγιση, η απόσταση υπολογίζεται ως εξής:

$$d(C_q, C_s) = max(d(C_i, C_s), d(C_j, C_s)).$$

3. Στην *weighted pair group* προσέγγιση,

$$d(C_q, C_s) = (d(C_i, C_s) + d(C_j, C_s))/2.$$

4. Στην *unweighted pair group* προσέγγιση,

$$d(C_q, C_s) = a_i \cdot d(C_i, C_s) + a_j \cdot d(C_j, C_s)),$$

όπου $a_i = \frac{|C_i|}{|C_i|+|C_j|}$ και $a_j = \frac{|C_j|}{|C_i|+|C_j|}$.

Αλγόριθμος 3 Ιεραρχική ομαδοποίηση

repeat

1. Βρίσκονται δύο ομάδες C_i και C_j όπου

$$d(C_i, C_j) = min_{r \neq s} d(C_r, C_s)$$

2. Συγχωνεύουμε τα C_i και C_j σε μία νέα ομάδα C_q.
3. Αντικαθίστανται τα C_i και C_j με την νέα ομάδα C_q.
4. Ανανεώνεται ο πίνακας αποστάσεων των νέων ομάδων.

until Όλα τα γονίδια να βρίσκονται στην ίδια ομάδα.

Κεφαλαιο 4

Bayesian Δικτυα

4.1 Εισαγωγή

Πολλές τεχνικές για εκπαίδευση βασίζονται κυρίως σε δεδομένα. Σε αντίθεση, η γνώση που περιέχεται σε έμπειρα συστήματα συνήθως προέρχεται από ειδικούς. Τα Bayesian δίκτυα συνδυάζουν τα καλύτερα στοιχεία και από τις δύο προσεγγίσεις. Πιο συγκεκριμένα, η αναπαράσταση κάποιου προβλήματος με Bayesian δίκτυα επιτρέπει να συνδυάσουμε στατιστικά δεδομένα με την υπάρχουσα γνώση.

Ένα Bayesian δίκτυο είναι μια γραφική αναπαράσταση της γνώσης. Επιπρόσθετα, η αναπαράσταση έχει τυπικές πιθανοτικές έννοιες, γεγονός που την κάνει βολική για στατιστική επεξεργασία [34]. Την τελευταία δεκαετία, τα Bayesian δίκτυα έχουν εξελιχθεί σε μια πολύ διαδεδομένη αναπαράσταση για να περιγράψουμε την αβεβαιότητα στην γνώση. Πιο πρόσφατα, οι ερευνητές ανέπτυξαν μεθόδους για την μάθηση Bayesian δικτύων με συνδυασμό εκ των προτέρων γνώσης και δεδομένων. Οι τεχνικές που αναπτύχθηκαν είναι καινούργιες και εξελισσόμενες, αλλά έχουν αποδειχθεί εντυπωσιακά αποτελεσματικές σε ορισμένα προβλήματα.

Χρησιμοποιώντας Bayesian δίκτυα, η διαδικασία εκπαίδευσης έχει ως εξής: Αρχικά εισάγεται η υπάρχουσα πληροφορία σε ένα Bayesian δίκτυο όπως κατά την κατασκευή έμπειρων συστημάτων. Στην συνέχεια, χρησιμοποιούνται τα δεδομένα για να την ενημερώση αυτής της γνώσης, δημιουργώντας ένα ή περισσότερα καινούργια Bayesian δίκτυα. Το αποτέλεσμα περιέχει έναν επαναπροσδιορισμό της αρχικής γνώσης και συνήθως την αναγνώριση νέων διαχωρισμών και συσχετίσεων. Η προσέγγιση αυτή είναι ανεκτική όσο αναφορά λάθη

της αρχική γνώσης. Ακόμα και όταν η αρχική γνώση δεν είναι πλήρης, αυτά τα στοιχεία μπορούν να χρησιμοποιηθούν για την βελτιώση της διαδικασίας μάθησης.

Η μάθηση με Bayesian δίκτυα είναι σε πολλά σημεία παρόμοια με αυτήν των νευρωνικών δικτύων. Η διαδικασία που εφαρμόζεται στα Bayesian δίκτυα, ωστόσο, έχει δύο σημαντικά πλεονεκτήματα. Πρώτον, μπορεί εύκολα να κωδικοποιηθεί η υπάρχουσα γνώση και να χρησιμοποιηθεί με σκοπό την αύξηση της αποτελεσματικότητας και της ακρίβειας στη μάθηση. Δεύτερον, οι κόμβοι και οι ακμές του δικτύου συνήθως αντιστοιχούν σε αναγνωρίσιμους διαχωρισμούς και αιτιατές σχέσεις. Συνεπώς, κάποιος μπορεί πιο εύκολα να μεταφράσει και να αντιληφθεί την γνώση που εμπεριέχει αυτή η αναπαράσταση.

4.1.1 Γενικοί Ορισμοί

Παρακάτω παρατίθενται μερικοί ορισμοί που θα φανούν χρήσιμοι στην συνέχεια.

Παραγοντικά Μοντέλα

Πρώτα εισάγεται η έννοια των παραγοντικών μοντέλων (Factored Models). Χρησιμοποιούνται τα κεφαλαία γράμματα X, Y, Z για ονόματα μεταβλητών και τα αντίστοιχα μικρά x,y,z για συγκεκριμένες τιμές που παίρνουν οι μεταβλητές. Σύνολα μεταβλητών δηλώνονται με έντονα γράμματα $\mathbf{X}$,$\mathbf{Y}$,$\mathbf{Z}$ και αντίστοιχα οι τιμές που παίρνουν οι μεταβλητές ενός συνόλου με μικρά έντονα γράμματα $\mathbf{x}$,$\mathbf{y}$,$\mathbf{z}$. Στην εκπαίδευση από δεδομένα το ενδιαφέρον έγκειται στο να βρούμε την καλύτερη εξήγηση μέσα από ένα σύνολο πιθανών εξηγήσεων. Αυτές οι πιθανές εξηγήσεις ορίζονται από ένα σύνολο υποθέσεων που λαμβάνονται υπόψη. Έστω μία κλάση μοντέλων $\mathfrak{M}$ τέτοια ώστε κάθε μοντέλο $M \in \mathfrak{M}$ να παραμετροποιείται από ένα διάνυσμα Θ^M. Κάθε επιλογή τιμών του Θ^M ορίζει μια κατανομή πιθανότητας $P(\cdot : M, \Theta^M)$ στα πιθανά σύνολα δεδομένων, όπου M δηλώνει την υπόθεση ότι η κατανομή που παράγει τα δεδομένα, προέρχεται από το μοντέλο M. Ένα παραγοντικό μοντέλο M για ένα σύνολο μεταβλητών $\mathcal{U} = \{X_1, ..., X_n\}$ είναι μια παραμετρική οικογένεια με παραμέτρους $\Theta^M = (\Theta_1^M, ..., \Theta_k^M)$ το οποίο ορίζει μία από κοινού κατανομή πιθανότητα της μορφής:

$$P(X_1, ..., X_n | M, \Theta^M) = \prod_i^n f_i^M(X_1, ..., X_n; \Theta_i^M), \tag{4.1}$$

όπου κάθε f_i^M είναι ένας παράγοντας του οποίου η τιμή εξαρτάται άπο κάποιες μεταβλητές $X_1, ..., X_n$. Ένα παραγοντικό μοντέλο είναι διαχωρίσιμο όταν ο χώρος των ικανών τιμών των παραμέτρων Θ^M προκύπτει από το εξωτερικό γινόμενο των ικανών τιμών των παραμέτρων Θ_i^M για κάθε f_i^M. Με άλλα λόγια, όταν μπορούν να συνδυστούν ικανές παραμετροποιήσεις διαφορετικών παραγόντων χωρίς περιορισμούς. Ικανές τιμές για τις παραμέτρους Θ^M είναι αυτές που ικανοποιούν τους περιορισμούς $\theta_{ijk}^M > 0$ και $\sum_k \theta_{ijk}^M = 1$ όπου k η ανάθεση τιμής για την μεταβλητή i, και j μία ανάθεση τιμών για τις μεταβλητές από τις οποίες εξαρτάται η μεταβλητή i.

Υπόθεση 4.1. Όλα τα μοντέλα M είναι διαχωρίσιμα παραγοντικά μοντέλα.

Υπόθεση 4.2. Για ένα μοντέλο M, κάθε μεταβλητή είναι ανεξάρτητη από τους προγόνους της, δοθέντος του συνόλου των πατέρων της.

$$P(X_1, ..., X_n|M, \Theta^M) = \prod_i^n f_i^M(X_i|Pa_i; \Theta_i^M) \tag{4.2}$$

Αυτή η συνθήκη ονομάζεται και *υπό συνθήκη ανεξαρτησία* ή συνθήκη *Markov* και είναι βασική υπόθεση στα Bayesian δίκτυα.

4.1.2 Ορισμός Bayesian δικτύων

Ένα Bayesian δίκτυο είναι ένα ακυκλικό κατευθυνόμενο γράφημα που αναπαριστά μια από κοινού κατανομή πιθανότητας στο σύνολο μεταβλητών $\mathfrak{U}$. Μπορεί να δηλωθεί από μια τριπλέτα $B = \langle G, P, \Theta \rangle$. Το πρώτο στοιχείο G, είναι ένα ακυκλικό κατευθυνόμενο γράφημα που οι κόμβοι αντιστοιχούν στις μεταβλητές $X_1, ..., X_n$ του $\mathfrak{U}$ και αναπαριστά το ακόλουθο σύνολο υποθέσεων ανεξαρτησίας: κάθε μεταβλητή X_i είναι ανεξάρτητη από όλες τις άλλες μεταβλητές δοθέντος του συνόλου των πατέρων της. Το δέυτερο στοιχείο P είναι ένα σύνολο από τοπικά μοντέλα $P_1, ..., P_n$. Κάθε τοπικό μοντέλο αντιστοιχίζει τις πιθανές τιμές Pa_i του συνόλου Pa_i πατέρων του X_i, σε ένα μέτρο πιθανότητας για το X_i. Τα τοπικά μοντέλα παραμετροποιούνται από τις παραμέτρους Θ_i. Ένα Bayesian δίκτυο B ορίζει μια μοναδική από κοινού κατανομή πιθανότητας P_B για το $\mathfrak{U}$ που δίνεται ως εξής:

$$P_B(X_1, ..., X_n) = \prod_{i=1}^n P_i(X_i|Pa_i; \Theta_i). \tag{4.3}$$

Είναι προφανές ότι τα Bayesian δίκτυα είναι παραγοντικά μοντέλα. Επιπλέον είναι διαχωρίσιμα από την στιγμή που ο συνδυασμός των ικανών τιμών για τις τοπικές παραμέτρους ορίζει ένα μέτρο πιθανότητας.

Η επόμενη υπόθεση είναι ως προς την επιλογή των παραγόντων στα παραγοντικά μοντέλα. Οι παράγοντες απαιτείται να ανήκουν στην εκθετική οικογένεια [35]. Ένας παράγοντας είναι εκθετικός αν και μόνο αν μπορεί να οριστεί με την μορφή

$$f(\mathbf{X} : \Theta) = e^{t(\Theta) \cdot s(\mathbf{X})} \tag{4.4}$$

όπου $t(\Theta)$ και $s(\mathbf{X})$ είναι συναρτήσεις με διανυσματικές τιμές ίδιας διάστασης και $(\cdot)$ το εσωτερικό γινόμενο. Οι εκθετικοί παράγοντες περιλαμβάνουν τις πολυωνυμικές (multinomial), μονοδιάστατες και πολυδιάστατες Gaussian κατανομές όπως και άλλες γνωστές κατανομές [35].

Υπόθεση 4.3. Όλα τα μοντέλα στο $\mathfrak{M}$ περιέχουν μόνο εκθετικούς παράγοντες [1].

Μια ακόμη βασική υπόθεση στα Bayesian δίκτυα είναι η ανεξαρτησία των δεδομένων.

[1] Δεν σημαίνει πως δεν μπορούμε να έχουμε παράγοντες που δεν είναι εκθετικοί. Όπως θα δούμε στην συνέχεια χρησιμοποιούμε και μη εκθετικούς παράγοντες. Σε πολλά όμως θεωρήματα και συμπεράσματα στην συνέχεια είναι απαραίτητη προϋπόθεση.

Υπόθεση 4.4. Τα στιγμιότυπα στο σύνολο δεδομένων D προκύπτουν ανεξάρτητα μεταξύ τους, δοθέντος ενός Bayesian δικτύου.

$$P(D|M) = \prod_{i=1}^{N} P(D_i|M) \tag{4.5}$$

Για να δείξουμε την αναπαράσταση, ας λάβουμε υπόψη το χώρο της επίλυσης του προβλήματος ενός αυτοκινήτου που δεν ξεκινά. Το πρώτο βήμα για την κατασκευή ενός Bayesian δικτύου είναι να αποφασίσουμε τις μεταβλητές και τις καταστάσεις του μοντέλου. Μια πιθανή επιλογή για τις μεταβλητές του προβλήματος είναι η *Μπαταρία (Μ)* με καταστάσεις *καλή* και *κακή*, τα *Καύσιμα (Κ)* με καταστάσεις *γεμάτο* και *άδειο*, ο *Μετρητής (ΜΤ)* με καταστάσεις *γεμάτο* και *άδειο*, *Γύρνα πίσω (Γ)* με καταστάσεις *ναι* και *όχι* και *Εκκίνηση (Ε)* με καταστάσεις *ναι* και *όχι*. Φυσικά θα μπορούσαμε να λάβουμε υπόψη και πολλές άλλες μεταβλητές όπως και στο φυσικό πρόβλημα, ή ακόμα να αναπαραστήσουμε κάποιες απο τις μεταβλητές με συνεχείς τιμές.

Το επόμενο βήμα στην κατασκευή ενός Bayesian δικτύου είναι η κατασκευή ενός κατευθυνόμενου ακυκλικού γραφήματος που θα κωδικοποιεί τις υποθέσεις μας για την υπό συνθήκη ανεξαρτησία των μεταβλητών. Δοθέντος του συνόλου $\mathcal{U} = \{X_1, ...X_n\}$, η από κοινού κατανομή του $\mathcal{U}$ μπορεί να γραφεί, χρησιμοποιώντας τον κανόνα της αλυσίδας των πιθανοτήτων ως εξής:

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i|X_1, ..., X_{i-1}) \tag{4.6}$$

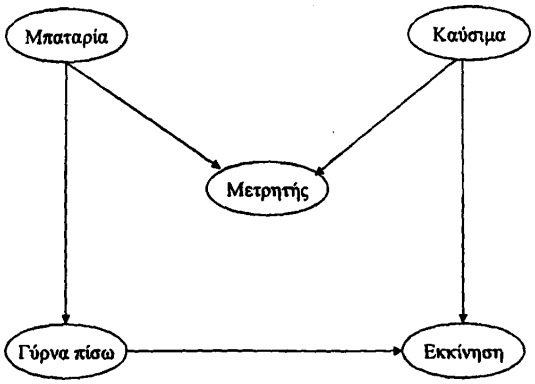
Τώρα για κάθε μεταβλητή X_i, θα υπάρχει κάποιο υποσύνολο $Pa_i \subseteq \{X_1, ...X_n\}$ τέτοιο ώστε τα X_i και $\{X_1, ...X_n\}$ να είναι υπό συνθήκη ανεξάρτητα μεταξύ τους δοθέντος του συνόλου Pa_i. Έτσι

$$P(X_i|X_1, ..., X_n) = P(X_i|Pa_i) \tag{4.7}$$

Οι υπό συνθήκη ανεξαρτησίες καθορίζουν την δομή του Bayesian δικτύου. Οι κόμβοι στην δομή αντιστοιχούν σε μεταβλητές στο πρόβλημα. Οι πατέρες του X_i αντιστοιχούν στο σύνολο Pa_i. Στο παράδειγμα μας, χρησιμοποιώντας την σειρά M, K, MT, Γ και E προκύπτουν οι υπό συνθήκη ανεξαρτησίες

$$\begin{aligned} P(K|M) &= P(K) \\ P(MT|M,K) &= P(MT|M,K) \\ P(\Gamma|M,K,MT) &= P(\Gamma|M) \\ P(E|M,K,MT,\Gamma) &= P(E|K,\Gamma) \end{aligned} \tag{4.8}$$

Συνεπώς παίρνουμε την δομή του Σχήματος (4.1).



Σχήμα 4.1: Ένα απλό δίκτυο για την επίλυση του προβλήματος για ένα αυτοκίνητο που δεν ξεκινά

4.2 Μάθηση των παραμέτρων του δικτύου για γνωστή δομή

Έστω ο χώρος μεταβλητών $\mathcal{U} = \{X_1, ..., X_n\}$ και ένα σύνολο δειγμάτων D που προέρχεται από την από κοινού κατανομή του χώρου. Γενικά για ένα τυχαίο δείγμα που προέρχεται από μια κατανομή πιθανότητας, οι εκ των προτέρων πιθανότητες για τις παραμέτρους της κατανομής μπορούν να ανανεωθούν. Αυτή η ανανέωση είναι σχετικά απλή (στην περίπτωση που τα δεδομένα είναι πλήρη) όταν οι εκ των προτέρων πιθανότητες ανήκουν στην οικογένεια των συζυγών κατανομών [35]. Πιο συγκεκριμένα αν υποτεθεί ότι ένα γνωστό δίκτυο M αναπαριστά την από κοινού κατανομή του χώρου και ότι οι παράμετροι είναι ανεξάρτητες μεταξύ τους, μπορούν να βρεθούν οι παραμέτροι που ταιριάζουν καλύτερα στα δεδομένα D και στις εκ των προτέρων πιθανοτήτες των παραμέτρων.

4.2.1 Διακριτές μεταβλητές- Πολυνυμικές (multinomial) κατανομές

Έστω η διακριτή μεταβλητή X_i με αριθμό πιθανών καταστάσεων r_i και παραμέτρους $\theta_{ijk} = P(X_i = k|Pa_i = j))$. Τα επαρκή στατιστικά (sufficient statistics) είναι τα N_{ijk} δηλαδή ο αριθμός των γεγονότων $X_i = k$ και $Pa_i = j$ όπου k μία πιθανή τιμή της μεταβλητής X_i και j μία πιθανή ανάθεση τιμών στο σύνολο πατέρων Pa_i της μεταβλητής X_i. Οι *Maximum Likelihood* (ML) παράμετροι δίνονται από την σχέση:

$$\widehat{\theta}_{ijk} = \frac{N_{ijk}}{N_{ij}} \tag{4.9}$$

όπου $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. Έστω τώρα ότι κάθε σύνολο παραμέτρων $\Theta_{ij} = \{\theta_{ij1}, ..., \theta_{ijx}\}$ ακολουθεί μια Dirichlet εκ των προτέρων κατανομή

$$P(\Theta_{ij}|M) = c \cdot \prod_{k}^{r_i} \theta_{ijk}^{a_{ijk}-1} \tag{4.10}$$

όπου c μια σταθερά κανονικοποίησης και a_{ijk} υπερπαράμετροι της Dirichlet κατανομής. Η εκ των υστέρων πιθανότητα των παραμέτρων δοθέντων του μοντέλου και των δεδομένων είναι:

$$P(\Theta_{ij}|D, M) = c' \cdot \prod_{k}^{r_i} \theta_{ijk}^{\alpha_{ijk}+N_{ijk}-1} \tag{4.11}$$

όπου c' μια άλλη σταθερά κανονικοποίησης. Οι *Maximum a posteriori* (MAP) παράμετροι είναι

$$\widehat{\theta}_{ijk} = \frac{N_{ijk} + \alpha_{ijk}}{N_{ij} + \alpha_{ij}} \tag{4.12}$$

όπου $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$.

4.2.2 Συνεχείς μεταβλητές- Gaussian κατανομή

Έστω η συνεχής μεταβλητή X_i και το σύνολο πατέρων της Pa_i. Η υπό συνθήκη κατανομή είναι

$$f(x_{ij}|pa_{ij}) = (2\pi\sigma_i^2)^{-\frac{1}{2}} exp\left(-\frac{1}{2\sigma_i^2}(x_{ij} - u_{ij})^2\right)$$

όπου $u_{ij} = \mu_i + \sum_{k \in Pa_i} b_k(x_{kj} - \mu_k)$, b_k οι συντελεστές παλινδρόμησης στις ακμές εισερχομένες στον κόμβο i από τους πατέρες του, μ_i, σ_i το μέσο και η διασπορά της μεταβλητής i και μ_k το μέσο του πατέρα k. Ισοδύναμα, μπορεί να γραφεί ως

$$x_{ij} = \mu_i + \sum_{k \in Pa_i} b_k(x_{kj} - \mu_k) + \sigma_i W_i \tag{4.13}$$

όπου $W_i \sim N(0, 1)$ μία τυχαία μεταβλητή λευκού θορύβου.

Η προσέγγιση που ακολουθείται είναι η μοντελοποίηση της από κοινού κατανομής κάθε μεταβλητής και των πατέρων της σαν μια MVG (minimum variance gaussian), ο υπολογισμος των επαρκών στατιστικών και η εύρεση των ML παραμέτρων της. Τα επαρκή στατιστικά για μια MVG κατανομή για N δείγματα είναι $s_N = \sum_{\ell=1}^{N} x_\ell$ και $Q_N = \sum_{\ell=1}^{N} x_\ell x_\ell^T$ όπου x_l διάνυσμα με τις l-ιοστές τιμές της μεταβλητής και των πατέρων της στα δεδομένα. Οι ML παράμετροι για την από κοινού κατανομή δίνονται από τις ακόλουθες σχέσεις:

$$\widehat{\mu} = \frac{1}{N} s_N = \frac{1}{N} \sum_{\ell=1}^{N} x_\ell \tag{4.14}$$

και

$$\widehat{\Sigma}_N = \frac{1}{N} Q_N - \widehat{\mu}\widehat{\mu}^T \tag{4.15}$$

αφού $Var[X] = E[XX^T] - (E[X]E[X]^T)$. Έστω τώρα X_1 το παιδί και X_2 οι πατέρες

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \quad \mu_X = \begin{pmatrix} \mu_{X_1} \\ \mu_{X_2} \end{pmatrix}, \quad \Sigma_X = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \tag{4.16}$$

Οι παράμετροι του κόμβου δίνονται από την υπό συνθήκη κατανομή του X_1 δοθέντος του X_2. Προκύπτει

$$\mu_{X_1|X_2} = E[X_1|X_2 = x_2] = \mu_{X_1} + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_{X_2}) \tag{4.17}$$

και

$$\Sigma_{X_1|X_2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \tag{4.18}$$

Συγκρίνοντας τις (4.13), (4.17) και (4.18) οι τοπικές παράμετροι για κάθε κόμβο δίνονται ως εξής

$$B = \Sigma_{12}\Sigma_{22}^{-1} \tag{4.19}$$

$$\mu = \mu_{X_1} - B\mu_{X_2} \tag{4.20}$$

$$\Sigma = \Sigma_{11} - B\Sigma_{21} \tag{4.21}$$

4.3 Μαθαίνοντας την δομή του δικτύου

Στην συνέχεια εξετάζουμε την περίπτωση που είμαστε αβέβαιοι όχι μόνο για τις παραμέτρους, αλλά και για την δομή του δικτύου. Η αβεβαιότητα αυτή μπορεί να εκφραστεί αναθέτοντας μια εκ των προτέρων πιθανότητα $P(M)$ σε κάθε πιθανή δομή $M \in \mathfrak{M}$ του δικτύου.

Έστω ότι $P(D|M)$ εκφράζει την υπόθεση ότι τα δεδομένα D είναι ένα τυχαίο δείγμα από την Bayesian δομή M. Από τον κανόνα του Bayes για την εκ των υστέρων πιθανότητα του δικτύου M δοθέντος του συνόλου δεδομένων D προκύπτει:

$$P(M|D) = c \cdot P(M)P(D|M) \tag{4.22}$$

όπου $c = 1/\sum_{M'} P(M')P(D|M')$ σταθερά κανονικοποίησης και M' όλα τα πιθανά μοντέλα. Η εκ των υστέρων πιθανότητα των παραμέτρων δοθέντος του δικτύου M και του συνόλου δεδομένων D δίνεται ως:

$$P(\Theta_M|D, M) = \frac{P(\Theta_M|M)P(D|\Theta_M, M)}{P(D|M)} \tag{4.23}$$

όπου

$$P(D|M) = \int P(D|\Theta_M, M)P(\Theta_M|M)d\Theta_M \tag{4.24}$$

είναι η *πιθανοφάνεια περιθωρίου* (marginal likelihood). Έστω κάποια υπόθεση h, ορίζουμε την πιθανότητα η υπόθεση h να είναι αληθής δοθέντος του συνόλου δεδομένων D,

«ζυγίζοντας» (averaging) όλα τα πιθανά μοντέλα και τις παραμέτρους του σύμφωνα με τους κανόνες των πιθανοτήτων:

$$P(h|D) = \sum_M P(M|D)P(h|D,M), \tag{4.25}$$

$$P(h|D,M) = \int P(h|\Theta_M, M)P(\Theta_M|D,M)d\Theta_M. \tag{4.26}$$

Για παράδειγμα, αν h η υπόθεση ότι επόμενη παρατήρηση είναι x_{N+1}, τότε προκύπτει

$$P(x_{N+1}|D) = \sum_M (P(M|D) \int P(x_{N+1}|\Theta_M, M)P(\Theta_M|D,M)d\Theta_M, \tag{4.27}$$

όπου $P(x_{N+1}|\Theta_M, M)$ η πιθανοφάνεια για το μοντέλο M. Η παραπάνω προσέγγιση αναφέρεται συνήθως ως *Bayesian «ζύγισμα» μοντέλων* (*Bayesian model averaging*). Παρατηρούμε ότι δεν μαθαίνεται μία μόνο δομή. Αντίθετα, «ζυγίζονται» όλα τα πιθανά μοντέλα από την εκ των υστέρων πιθανοφάνεια τους. Η μέθοδος αυτή δεν είναι κατάλληλη για κάθε είδους ανάλυση. Για παράδειγμα, μπορεί να επιθυμούμε μόνο ένα ή μερικά μοντέλα για την κατανόηση του προβλήματος ή για γρήγορες προβλέψεις. Σε αυτήν την περίπτωση, επιλέγουμε ένα ή μερικά «καλά» μοντέλα από όλα τα πιθανά. Η παραπάνω προσέγγιση ονομάζεται *επιλογή μοντέλου* (*model selection*) όταν γίνεται επιλογή ενός μοναδικού μοντέλου και *επιλεκτικό «ζύγισμα» μοντέλων* (*selective model averaging*) όταν γίνεται επιλογή παραπάνω του ενός μοντέλου. Οι παραπάνω προσεγγίσεις είναι επίσης χρήσιμες όταν δεν είναι εφικτό να «ζυγιστούν» όλα τα πιθανά μοντέλα.

4.3.1 Μέτρα Αξιολόγησης (Scoring Metrics)

Το πιο σημαντικό ερώτημα είναι κατά πόσο μπορεί να προσεγγιστεί η πιθανότητα $P(x_{N+1}|D)$ χρησιμοποιώντας ένα μικρό σχετικά σύνολο μοντέλων. Αυτό το ερώτημα είναι δύσκολο να απαντηθεί θεωρητικά. Ωστόσο, πολλοί ερευνητές έδειξαν πειραματικά ότι και μόνο μία «καλή» δομή δικτύου συχνά μπορεί να δώσει μια πολύ καλή προσέγγιση [36]. Αυτό το συμπέρασμα είναι και η βασική αιτία για το μεγάλο ερευνητικό ενδιαφέρον που υπάρχει σχετικά με τα Bayesian δίκτυα τα τελευταία χρόνια.

Με δεδομένο το προηγούμενο συμπέρασμα, ένα άλλο σημαντικό θέμα είναι ο προσδιορισμός του πόσο "καλή" είναι μια δομή δικτύου. Η προσέγγιση που κυρίως χρησιμοποιείται είναι ο συνδυασμός ενός μέτρου αξιολόγησης και ενός ευρετικού αλγορίθμου. Τα μέτρα αξιολόγησης παίρνουν μια εκ των προτέρων γνώση, ένα σύνολο δεδομένων και ένα σύνολο δομών του δικτύου και υπολογίζουν πόσο καλά ταιριάζει κάθε δίκτυο στα δεδομένα και στην εκ των προτέρων γνώση. Οι ευρετικοί αλγόριθμοι μας δίνουν τις δομές που πρόκειται να αξιολογηθούν.

Ένα προφανές μέτρο για μία δομή είναι η σχετική εκ των υστέρων πιθανότητα του δικτύου δοθέντων των δεδομένων. Για παράδειγμα, μπορούμε να γράψουμε $P(D,M) = P(M)P(D|M)$ ή να υπολογίσουμε τον παράγοντα Bayes $P(D|M)/P(D|M_{S0})$ όπου M_{S0} είναι ένα δίκτυο αναφοράς όπως το κενό δίκτυο. Στην συνέχεια περιγράφουμε τις σημαντικότερες προσεγγίσεις της πιθανότητας των δεδομένων δοθέντος του δικτύου M.

Bayesian Dirichlet (BD) μέτρο

Υπό ορισμένες υποθέσεις οι υπολογισμοί που απαιτούνται για τα Bayesian «ζύγισμα» μοντέλων, επιλογή μοντέλου, επιλεκτικό «ζύγισμα» μοντέλων μπορούν να γίνουν αποτελεσματικά και σε κλειστή μορφή. Οι υποθέσεις αυτές είναι:

1. Κάθε μεταβλητή είναι διακριτή. Χρησιμοποιούμε x_{ik} και pa_{ij} για να δηλώσουμε την k-οστή πιθανή κατάσταση της μεταβλητής X_i και την j-οστή πιθανή ανάθεση τιμών του συνόλου πατέρων Pa_i στο μοντέλο, της μεταβλητής X_i. Επίσης, χρησιμοποιούμε r_i και q_i για να δηλώσουμε τον αριθμό των πιθανών καταστάσεων της X_i και τον αριθμό των πιθανών αναθέσεων τιμών του συνόλου Pa_i αντίστοιχα.

2. Κάθε τοπική κατανομή $P(x_{ik}|pa_{ij}, \Theta_M, M)$ αποτελείται από ένα σύνολο από πολυωνυμικές κατανομές, μία πολυωνυμική κατανομή για κάθε i και j. Έτσι

$$P(x_{ik}|pa_{ij}, \Theta_m, M) = \theta_{ijk},$$

όπου θ_{ijk} παράμετροι που ικανοποιούν $\theta_{ijk} > 0$ για κάθε i, j και k και $\sum_{k=1}^{r_i} \theta_{ijk} = 1$ για κάθε i και j. Για ευκολία εισάγεται το σύνολο παραμέτρων

$$\Theta_{ij} = \{\theta_{ij1}, ..., \theta_{ijk}\},$$

για κάθε i και j.

3. Τα σύνολα παραμέτρων Θ_{ij} είναι αμοιβαία ανεξάρτητα:

$$P(\Theta_M|M) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} P(\Theta_{ij}|M).$$

4. Κάθε σύνολο παράμετρων Θ_{ij} έχει μια Dirichlet κατανομή:

$$P(\Theta_{ij}|M) = Dir(\Theta_{ij}|a_{ij1}, ..., a_{ijr_i}) \propto \prod_{k=1}^{r_i} \theta_{ijk}^{a_{ijk}-1},$$

με υπερ-παραμέτρους $a_{ijk} > 0$ για κάθε i, j και k.

5. Το σύνολο δεδομένων D είναι πλήρες.

Υπό αυτές τις συνθήκες, οι παράμετροι παραμένουν ανεξάρτητοι δοθέντος ενός τυχαίου δείγματος του D που δεν περιέχει χαμένη παρατήρηση,

$$P(\Theta_M|D, M) = \prod_{i=1}^{n} \prod_{j=1}^{r_i} P(\Theta_{ij}|D, M), \tag{4.28}$$

και η εκ των υστέρων κατανομή πιθανοφάνειας για κάθε Θ_{ij} ακολουθεί μια Dirichlet κατανομή

$$P(\Theta_{ij}|D, M) = Dir(\Theta_{ij}|a_{ij1} + N_{ij1}, ..., a_{ijr_i} + N_{ijr_i}), \tag{4.29}$$

όπου N_{ijk} ο αριθμός των περιπτώσεων στο D όπου $X_i = x_{ik}$ και $Pa_i = pa_{ij}$. Επιπρόσθετα, παίρνουμε για την περιθωριοποιημένη πιθανοφάνεια

$$P(D|M) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij} + N_{ij})} \cdot \prod_{k=1}^{r_i} \frac{\Gamma(N'_{ijk} + N_{ijk})}{\Gamma(N'_{ijk})}, \tag{4.30}$$

όπου $a_{ij} = \sum_{k=1}^{r_i} a_{ijk}$ και $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. Η απόδειξη δίνεται στην αναφορά [36].

BIC και MDL μέτρο

Μια άλλη προσέγγιση, πολύ διαδεδομένη, είναι η προσέγγιση της εκ των υστέρων πιθανότητας. Υποτίθεται ότι η πιθανότητα συγκεντρώνεται γύρω από την περιοχή των παραμέτρων $\widehat{\Theta}_M$ που μεγιστοποιούν την εκ των υστέρων πιθανότητα. Οπότε χρησιμοποιείται η προσέγγιση $P(D|M) \approx P(D|M, \widehat{\Theta}_M)$. Όμως δεν μπορεί να χρησιμοποιηθεί αυτούσια αυτή η ποσότητα επειδή ευνοεί τα πιο πολύπλοκα δίκτυα. Αυτό οφείλεται στο γεγονός ότι δεν τίθενται ισχυροί περιορισμοί στις παραμέτρους Θ_M. Με λίγα λόγια παρατηρείται το φαινόμενο της υπερεκπαίδευσης (overfitting). Για αυτό ακριβώς τον λόγο χρησιμοποιείται και ένα κριτήριο το όποιο δίνει μεγάλυτερο βάρος στα πιο απλά δίκτυα σε σχέση με τα πιο πολύπλοκα. Ο Akaike [37] πρότεινε το έξης μέτρο αξιολόγησης

$$\log P(D|M, \widehat{\Theta}_M) + Dim(M) \tag{4.31}$$

όπου M το μοντέλο και $Dim(M)$ ο αριθμός των λογικά ανεξάρτητων παραμέτρων του M. Αυτό το μέτρο αξιολόγησης ονομάζεται Akaike information criterion (AIC). Για ένα Bayesian δίκτυο, έχουμε

$$Dim(M) = \prod_{i=1}^{n} q_i(r_i - 1) \tag{4.32}$$

Ο Schwarz [38] πρότεινε ένα παρόμοιο μέτρο αξιολόγησης με όρο ποινής $PEN = \frac{1}{2} Dim(M) \log(N)$ το οποίο είναι γνωστό ως Bayesian Information criterion (BIC). Ένα παρόμοιο με το BIC μέτρο είναι και το minimum description length (MDL) όπου σαν ποινή ορίζεται ο αριθμός των bit που χρειάζονται για να κωδικοποιηθούν τα δεδομένα δοθέντος του μοντέλου.

Μια σημαντική παρατήρηση είναι ότι καθώς ο αριθμός των δεδομένων D πλησιάζει θεωρητικά το άπειρο, το BD με ομοιόμορφη (uniform) εκ των προτέρων πιθανότητα στις παραμέτρους, τα BIC και MDL μας δίνουν το ίδιο μέτρο. Δυστυχώς αυτό στην πράξη δεν είναι εφικτό.

4.3.2 Εκ των προτέρων πιθανότητες στην δομή

Η εκ των υστέρων πιθανότητα και αρκετά μέτρα αξιολόγησης στη θεωρία απόφασης απαιτούν την ανάθεση εκ των προτέρων πιθανοτήτων σε όλες τις πιθανές δομές. Παρακάτω περιγράφουμε ένα τρόπο που αυτό επιτυγχάνεται, όπως δόθηκε από τους Heckerman *et al.* [39].

Η προσέγγιση αυτή απαιτεί ο χρήστης να φτιάξει την δομή ενός εκ των προτέρων δικτύου για το πρόβλημα. Η μέθοδος υποθέτει ότι αυτό το δίκτυο είναι η «καλύτερη» υπόθεση του χρήστη για την δομή του δικτύου που αναπαριστά τις φυσικές πιθανότητες. Δοθείσας μιας εκ των προτέρων δομής του δικτύου P, υπολογίζουμε τις εκ των προτέρων πιθανότητες των δικτύων M ως εξής. Για κάθε μεταβλητή X_i στο U, έστω δ_i δηλώνει τον αριθμό των κόμβων στην συμμετρική διαφορά του $Pa_i(M)$ και $Pa_i(P)$, που είναι $Pa_i(M \bigcup Pa_i(P)) \backslash (M \bigcap Pa_i(P))$. Τότε, το M και το εκ των προτέρων δίκτυο διαφέρουν συνολικά κατά $\delta = \sum_{i=1}^{n} \delta_i$ ακμές. Υπολογίζουμε την εκ των προτέρων πιθανότητα θέτοντας σαν ποινή για το M μια σταθερά $0 < \kappa \leq 1$ για κάθε τέτοια ακμή. Έτσι τίθεται

$$P(M) = \gamma \kappa^{\delta} \tag{4.33}$$

όπου γ μια σταθερά κανονικοποίησης, που μπορεί να αγνοηθεί.

Θέτοντας διαφορετικές ποινές για κάθε κόμβο X_i και για διαφορετικά σύνολα πατέρων για κάθε κόμβο, δίνεται περισσότερη πληροφορία. Το πόσο μεγάλη είναι η ποινή για κάθε κόμβο, δηλώνει το πόσο ισχυρή είναι η πεποίθηση μας για την γειτονιά του κόμβου. Μια άλλη περίπτωση είναι να τεθεί κατηγορηματικά ότι κάποιες ακμές πρέπει να εμφανίζονται στην δομή. Στην εξίσωση (4.33) τίθεται μηδενική η εκ των προτέρων πιθανότητα σε δίκτυα που δεν περιέχουν τους περιορισμούς που έχουν οριστεί.

4.3.3 Εκ των προτέρων πιθανότητες στις παραμέτρους

Όπως και στις δομές πρέπει να οριστούν εκ των προτέρων πιθανότητες και για τις παραμέτρους. Αυτή η πληροφορία μπορεί να χρησιμοποιηθεί για να υπολογίσουμε τα AIC και BIC κριτήρια πιο αποτελεσματικά. Πολλοί ερευνητές έχουν προτείνει παρόμοιες προσεγγίσεις όταν έχουμε πολλά πιθανά δίκτυα [36,40]. Η προσέγγιση που ακολουθείται είναι βασισμένη στα αποτελέσματα των Heckerman και Geiger [41]. Αν όλες οι επιτρεπτές τιμές των φυσικών πιθανοτήτων είναι πιθανές, τότε η ανεξαρτησία των παραμέτρων και η ισοδυναμία των υποθέσεων, απαιτεί οι φυσικές πιθανότητες για πλήρες δομές δικτύων να έχουν Dirichlet κατανομές όπως ορίζονται στην εξίσωση (4.10). Επιπλέον θα πρέπει να ισχύει ο περιορισμός

$$a_{ijk} = N' P(x_i = k, Pa_i = j | M_{Sc}), \tag{4.34}$$

όπου N' είναι το ισοδύναμο μέγεθος δείγματος του χρήστη, M_{Sc} ένα οποιοδήποτε πλήρες δίκτυο και $P(x_i = k, Pa_i = j | M_{Sc})$ είναι η πιθανότητα του χρήστη για $x_i = k$ και $Pa_i = j$ για την πρώτη περίπτωση στα δεδομένα. Υπό αυτές τις συνθήκες, για ένα πλήρες δίκτυο, οι εκ των προτέρων πιθανότητες στις παραμέτρους ορίζονται

1. Από την κατασκεύη του εκ των προτέρων δικτύου.

2. Από την υπόθεση για το μέγεθος του ισοδύναμου δείγματος στο εκ των προτέρων δίκτυο.

Για να καθοριστούν οι εκ των προτέρων παράμετροι για μη πλήρεις δομές οι Heckerman *et al.* [39] κάνουν την εξής υπόθεση:

Υπόθεση 4.5. Αν σε δύο δομές M_{S1} και M_{S2}, η μεταβλητή X_i έχει τους ίδιους πατέρες στα M_{S1} και M_{S2}, τότε

$$P(\Theta_{ij}|M_{S1}) = P(\Theta_{ij}|M_{S2}) \tag{4.35}$$

για $j = 1, ..., q_i$.

Αυτή η ιδιότητα αποκαλείται τοπικότητα παραμέτρων, επειδή μας λέει ότι οι κατανομές με παραμέτρους Θ_{ij} εξαρτώνται μόνο από την τοπική δομή του δικτύου στην μεταβλητή X_i, δηλαδή από την μεταβλητή X_i και τους πατέρες της.

Κάνοντας τις υποθέσεις για ανεξαρτησία και τοπικότητα των παραμέτρων, είναι απλό να κατασκευαστούν οι εκ των προτέρων πιθανότητες για ένα πλήρες δίκτυο. Συγκεκριμένα, δεδομένου της ανεξαρτησίας των παραμέτρων, κατασκευάζονται οι εκ των προτέρων πιθανότητες για κάθε κόμβο ξεχωριστά. Επιπλέον, αν ο κόμβος X_i έχει πατέρες Pa_i, στο δοθέν δίκτυο, είναι δυνατό να αναγνωριστεί ένα πλήρες δίκτυο όπου το X_i έχει αυτούς τους πατέρες και να χρησιμοποιηθεί η τοπικότητα των παραμέτρων ώστε να καθοριστούν οι εκ των προτέρων πιθανότητες για αυτόν τον κόμβο. Το αποτέλεσμα είναι μια ειδική περίπτωση του BD, που ονομάζεται BDe (το 'e' προκύπτει από το ισοδύναμο ως προς την πιθανοφάνεια-equivalent), και αναθέτει ίδιο μέτρο σε ισοδύναμες δομές του δικτύου.

4.3.4 Ευρετικοί Αλγόριθμοι

Σε αυτή τη παράγραφο περιγράφουμε ευρετικούς αλγορίθμους για αναγνώριση δικτύων με μεγάλο μέτρο αξιολόγησης. Όλοι οι αλγόριθμοι που θα περιγράψουμε βασίζονται στην *παραγοντική* ιδιότητα.

Ορισμός 4.1. Δοθείσας μιας δομής δικτύου για ένα χώρο $\mathcal{U}$, λέμε ότι το μέτρο αξιολόγησης της δομής είναι παραγοντικό αν μπορεί να γραφεί σαν ένας παράγοντας μέτρων, όπου το κάθε μέτρο είναι μια συνάρτηση αποκλειστικά του κόμβου και των πατέρων του.

Για παράδειγμα στην εξίσωση (4.30), η πιθανότητα $P(D|M)$ που δίνεται από το BD μέτρο είναι παραγοντική. Συνεπώς αν και οι εκ των προτέρων πιθανότητες είναι παραγοντικές είναι και το BD μέτρο παραγοντικό. Οπότε μπορούμε να γράψουμε

$$P(D, M) = \prod_{i=1}^{n} s(X_i|Pa_i) \tag{4.36}$$

όπου $s(X_i|Pa_i)$ είναι συνάρτηση του X_i και των πατέρων του. Τα περισσότερα Bayesian και μη Bayesian μέτρα είναι παραγοντικά. Δοθέντος ενός παραγοντικού μέτρου, μπορούν να συγκριθούν τα μέτρα δύο δικτύων που διαφέρουν κατά μία πρόσθεση ή αφαίρεση ακμής που δείχνει στο X_i, υπολογίζοντας μόνο τον όρο $s(X_i|Pa_i)$ και για τις δύο δομές.

Καταρχάς, ας εξετάσουμε την ειδική περίπτωση όπου αναζητείται το δίκτυο με το μεγαλύτερο μέτρο, ανάμεσα σε όλες τις δομές στις οποίες κάθε κόμβος έχει το πολύ έναν πατέρα. Για κάθε ακμή $X_j \rightarrow X_i$, ανατίθεται ένα βάρος $w(X_i, X_j) \equiv \log s(X_i|X_j) - \log(X_i|\emptyset)$. Από

την εξίσωση (4.36) , προκύπτει

$$\begin{aligned} \log P(D,M) &= \sum_{i=1}^{n} \log s(X_i|Pa_i) \\ &= \sum_{i=1}^{n} w(X_i,Pa_i) + \sum_{i=1}^{n} \log s(X_i,\emptyset), \end{aligned} \tag{4.37}$$

όπου Pa_i οι πατέρες του X_i. Ο δεύτερος όρος στην εξίσωση (4.37) είναι ίδιος για όλες τις δομές. Έτσι μεταξύ των δομών στις οποίες κάθε κόμβος έχει το πολύ έναν πατέρα, προκύπτει το ίδιο αποτέλεσμα, είτε βαθμολογώντας τα δίκτυα με το άθροισμα των βαρών $\sum_{i=1}^{n} w(X_i,Pa_i)$ είτε με το μέτρο της πιθανοφάνειας. Τώρα ας υποθέσουμε ότι θέλουμε να βρούμε το "καλύτερο" δίκτυο από το σύνολο δικτύων όπου κάθε κόμβος δεν έχει παραπάνω από k πατέρες. Δυστυχώς, το πρόβλημα για $k > 1$ είναι NP-δύσκολο (NP-hard) [42]. Οπότε στρεφόμαστε σε ευρετικούς αλγορίθμους.

Οι περισσότερες από τις ευρετικές μεθόδους για την μάθηση της δομής ενός Bayesian δικτύου, κάνουν διαδοχικές αλλαγές στις ακμές του δικτύου και χρησιμοποιούν την παραγοντική ιδιότητα για να αξιολογήσουν αυτήν την αλλαγή. Οι πιθανές αλλαγές είναι εύκολο να αναγνωριστούν. Για κάθε ζευγάρι μεταβλητών, εάν υπάρχει μεταξύ τους ακμή, αυτή μπορεί να αφαιρεθεί ή να αντιστραφεί. Εάν δεν υπάρχει μεταξύ τους ακμή, μπορεί να προστεθεί μια τέτοια ακμή. Όλες οι αλλαγές που γίνονται δεν πρέπει να δημιουργούν κατευθυνόμενο κύκλο στο γράφημα. Δηλώνονται με E οι πιθανές αλλαγές σε ένα δίκτυο και $\Delta(e)$ η διαφορά στο μέτρο του δικτύου από την αλλαγή $e \in E$. Δοθέντος ένος παραγοντικού μέτρου, εάν μία ακμή στο X_i προστεθεί η αφαιρεθεί, αρκεί να υπολογιστεί μόνο ο όρος $s(X_i|Pa_i)$ για να καθορίσουμε το $\Delta(e)$. Αν η μία ακμή μεταξύ των X_i και X_j αντιστραφεί τότε μόνο οι όροι $s(X_i|Pa_i)$ και $s(X_j|Pa_j)$ αρκεί να υπολογιστούν.

Αλγόριθμος 4 Αλγόριθμος Local Search

$i := 0$. Επιλέγεται ένα αρχικό γράφημα.

repeat

1. Υπολογίζεται το $\Delta(e)$ για κάθε $e \in E$
2. Πραγματοποιείται η αλλαγή e με το μεγαλύτερο $\Delta(e)$, εφόσον είναι θετικό

until Να μην υπάρχει θετικό $\Delta(e)$.

Ένας απλός ευρετικός αλγόριθμος αναζήτησης είναι ο *local search* [43], που περιγράφεται από τον Αλγόριθμο 4. Με την χρήση παραγοντικών μέτρων αξιολόγησης, αποφεύγεται ο υπολογισμός όλων των $\Delta(e)$ μετά από κάθε αλλαγή. Συγκεκριμένα, αν κανένα από τα X_i, X_j και τους πατέρες τους δεν έχει αλλάξει, το $\Delta(e)$ παραμένει το ίδιο για όλες τις αλλαγές e σε αυτούς τους κόμβους όσο το δίκτυο είναι ακυκλικό. Επιλογές για αρχικό δίκτυο είναι το άδειο δίκτυο, ένα τυχαίο δίκτυο και το εκ των προτέρων δίκτυο. Ένα πιθανό πρόβλημα για οποιαδήποτε τοπική μέθοδο αναζήτησης είναι ο εγκλωβισμός σε ένα τοπικό μέγιστο. Πιθανές μεθόδοι για την αποφυγή μιας τέτοιας κατάστασης είναι ο επαναληπτικός hill-climbing και simulated annealing μεθόδοι. Στο επαναληπτικό hill-climbing, εφαρμό-

ζεται τοπική αναζήτηση μέχρι την εύρεση ενός τοπικού μεγίστου. Τότε, τυχαία αλλάζει η τρέχουσα δομή του δικτύου, όπου επαναλαμβάνεται αυτή η διαδικασία αρκετές φορές.

Σε μία εκδοχή του simulated annealing αρχικοποιείται το σύστημα σε μια θερμοκρασία T_0. Στην συνέχεια επιλέγεται κάποια αλλαγή τυχαία και εκτιμάται η έκφραση $p = \exp(\Delta(e)/T_0)$. Αν $p > 1$ τότε πραγματοποιείται η αλλαγή e, αλλιώς εφαρμόζεται η αλλαγή με πιθανότητα p. Η διαδικασία επιλογής και εκτίμησης επαναλαμβάνεται α φορές ή μέχρι να συμβούν β αλλαγές. Αν δεν έχει προκύψει κάποια αλλαγή στις α επαναλήψεις, ο αλγόριθμος τερματίζεται. Αλλιώς μειώνεται η θερμοκρασία πολλαπλασιαζόμενη με κάποια σταθερά $0 < \gamma \leq 1$ και συνεχίζεται η διαδικασία αναζήτησης. Η αναζήτηση σταματάει την αναζήτηση όταν η θερμοκρασία χαμηλώσει πάνω από δ φορές. Έτσι ο αλγόριθμος ελέγχεται από πέντε παραμέτρους: T_0, α, β, γ και δ. Συνήθως ο αλγόριθμος αρχικοποιείται με άδειο δίκτυο και με αρκετά μεγάλο T_0, οπότε αρχικά κάθε επιτρεπτή αλλαγή πραγματοποιείται, δημιουργώντας έτσι ένα τυχαίο δίκτυο. Το μειονέκτημα των στοχαστικών μεθόδων όπως αυτή του simulated annealing είναι ότι απαιτούνται πολλές επαναλήψεις για να βρούμε ένα "καλό" δίκτυο. Έτσι ο αριθμός των απαραίτητων επαναλήψεων αυξάνει δραματικά με την αύξηση του αριθμού των κόμβων και συνεπώς και των πιθανών δικτύων. Αντίθετα, έχει αποδειχθεί ότι οι hill-climbing αλγόριθμοι δίνουν πολύ καλά αποτελέσματα και μία παράλληλη έκδοση αυτού του αλγορίθμου χρησιμοποιούμε και εμείς.

Κεφαλαιο 5

Bayesian Δικτυα και Ελλιπεις Τιμες

Σε αυτό το κεφάλαιο εξετάζουμε την περίπτωση που τα δεδομένα μας έχουν ελλιπείς τιμές. Υποθέτουμε ότι δεν έχουμε χρησιμοποιήσει μεθόδους ανάκτησης ελλιπών τιμών όπως αυτές αναφέρθηκαν στην παράγραφο 3.2. Εξετάζουμε μεθόδους που χειρίζονται τις ελλιπείς τιμές στη διαδικασία μάθησης των Bayesian δικτύων. Όταν οι παρατηρήσεις για μερικές μεταβλητές λείπουν, οι παράμετροι παύουν να είναι πλέον ανεξάρτητες μεταξύ τους και δεν μπορούν να χρησιμοποιηθούν οι κλειστοί τύποι, που αναφέραμε στο προηγούμενο κεφάλαιο, για να καθορισθεί η πιθανοφάνεια περιθωρίου. Μερικές προσεγγίσεις για τον υπολογισμό της πιθανοφάνειας περιθωρίου σε αυτήν την περίπτωση είναι οι προσεγγίσεις Monte Carlo (Gibbs δειγματοληψία, δειγματοληψία σημαντικότητας (Importance sampling)) και οι προσεγγίσεις μεγάλου δείγματος (Laplace προσέγγιση) [44]. Οι Monte Carlo προσεγγίσεις είναι ακριβείς αλλά συνήθως δεν είναι αποδοτικές, ενώ οι προσεγγίσεις μεγάλου δείγματος είναι πιο αποδοτικές από άποψη υπολογιστικού κόστους αλλά είναι ακριβείς μόνο όταν έχουμε μεγάλο σύνολο δεδομένων. Είναι σημαντικό να σημειώσουμε ότι έχει σημασία όταν χειριζόμαστε ελλιπή δεδομένα να λαμβάνουμε υπόψη την διαδικασία με την οποία αυτά λείπουν. Για παράδειγμα, για ένα χαμένο δεδομένο σε μία μελέτη φαρμάκων δεν μπορεί να αγνοηθεί η πιθανότητα, σαν αποτέλεσμα του φαρμάκου, ο ασθενής να είναι πολύ άρρωστος ώστε να λάβουμε τη μέτρηση. Σε αντίθεση αν τα δεδομένα λείπουν λόγω εργαστηριακών λαθών, είναι λογικό να αγνοήσουμε αυτό το αίτιο αφού θεωρείται τυχαίο. Όταν υπάρχει μια διαδικασία εξαιτίας της οποίας τα δεδομένα λείπουν, το μοντέλο (ή τα μοντέλα) θα πρέπει να επεκταθεί(ούν) ώστε να αναπαριστά και αυτήν την διαδικασία. Μία απλή προσέγγιση είναι η πρόσθεση δυαδικών μεταβλητών $(I_1..., I_N)$, όπου I_i η μεταβλητή που δείχνει αν η

μεταβλητή X_i λείπει ή όχι. Ο Rubin [45] μελετά αυτές τις διαδικασίες και μεθόδους για τον χειρισμό τους.

Το βασικό πρόβλημα των προσεγγίσεων Monte-Carlo και μεγάλου δείγματος, είναι το υπολογιστικό κόστος για την πρώτη και η απαίτηση για μεγάλο αριθμό δεδομένων για την δεύτερη. Οπότε για μεγάλα δίκτυα που είναι και το πρόβλημα που αντιμετωπίζουμε, η πρώτη προσέγγιση είναι απαγορευτική από άποψη υπολογιστικού κόστους. Επειδή όπως αναφέραμε τα δεδομένα μας δεν είναι επαρκή, τουλάχιστον στον βαθμό που θα καθιστούσε τις προσεγγίσεις μεγάλου δείγματος ικανοποιητικές δεν μπορούμε να αρκεστούμε ούτε σε αυτές. Ωστόσο θα τις παρουσιάσουμε παρακάτω.

5.1 Monte Carlo

Μία Monte-Carlo μέθοδος, που παρουσιάστηκε από τους Geman, Geman [46] είναι η Gibbs δειγματοληψία. Δοθέντος ενός συνόλου μεταβλητών $\mathbf{X} = (X_1, ..., X_n)$ και κάποιας από κοινού κατανομή πιθανότητας $p(\mathbf{x})$, μπορούμε να χρησιμοποιήσουμε ένα Gibbs δειγματολήπτη για να προσεγγιστεί η αναμενόμενη τιμή μιας συνάρτησης $f(\mathbf{x})$ ως προς την $p(\mathbf{x})$. Πρώτα, επιλέγεται μία αρχική κατάσταση των μεταβλητών στο σύνολο $\mathbf{X}$ με κάποιον τρόπο (π.χ. τυχαία). Στην συνέχεια, αναιρείται η τρέχουσα κατάσταση της X_1, και υπολογίζεται η κατανομή πιθανότητάς της, δοθέντος των καταστάσεων των υπολοίπων $n-1$ μεταβλητών. Επαναλαμβάνεται αυτή η διαδικασία για κάθε μεταβλητή $X_2, ..., X_n$,'δημιουργώντας έτσι ένα νέο δείγμα $\mathbf{x}$. Μετά η διαδικασία ανατρέχει στο προηγούμενο βήμα, κρατώντας τον μέσο όρο της $f(\mathbf{x})$ πάνω στα δείγματα που κατασκευάζονται. Μετά από κάποιο σχετικά μικρό αριθμό επαναλήψεων - γνωστή ως φάση «burn in»- οι πιθανές τιμές του $\mathbf{x}$ θα προκύπτουν με πιθανότητα $p(\mathbf{x})$. Συνεπώς, ο μέσος όρος της $f(\mathbf{x})$ πάνω σε αυτά τα δείγματα θα συγκλίνει στην $E_{p(\mathbf{x})}(f(\mathbf{x}))$. Ο τρόπος που υπολογίζεται η πιθανοφάνεια περιθωρίου είναι γνωστός ως μέθοδος *Υποψηφίου* (Candidate method). Αυτή η προσέγγιση είναι βασισμένη στον κανόνα του Bayes

$$P(D|M) = \frac{P(D|\theta^*_M, M)P(\theta^*_M|M)}{P(\theta^*_M|D, M)} \tag{5.1}$$

για οποιαδήποτε τιμή του θ^*_M. Για να υπολογιστεί η πιθανότητα $P(D|M)$, επιλέγεται κάποια τιμή θ^*_M (π.χ τυχαία), υπολογίζεται ο αριθμητής ακριβώς και προσεγγίζεται ο παρονομαστής με έναν Gibbs δειγματολήπτη. Για να προσεγγιστεί η πιθανότητα $P(\theta^*_M|D, M)$, πρώτα αρχικοποιούνται οι τιμές των μη παρατηρημένων μεταβλητών για κάθε περίπτωση με κάποιο τρόπο (π.χ. τυχαία). Σαν αποτέλεσμα, έχουμε ένα πλήρες σύνολο δεδομένων D_c. Στην συνέχεια, επιλέγεται κάποια τιμή X_{il} (μεταβλητή X_i στην περίπτωση l) που δεν είναι παρατηρημένη στο αρχικό δείγμα D, και επανυπολογίζεται η κατάσταση της σύμφωνα με την κατανομή πιθανότητας

$$P(x'_{il}|D_c\backslash x_{il}, M) = \frac{P(x'_{il}|D_c\backslash x_{il}|M)}{\sum_{x''_{il}} P(x''_{il}|D_c\backslash x_{il}|M)}, \tag{5.2}$$

όπου $D_c\backslash x_{il}$ δηλώνει το σύνολο D_c όπου η παρατήρηση x_{il} έχει αφαιρεθεί και το άθροισμα στον παρονομαστή είναι πάνω σε όλες τις πιθανές καταστάσεις της μεταβλητής X_{il}. Οι

όροι στον αριθμητή και στον παρονομαστή είναι πιθανοφάνειες περιθωρίου πάνω στο πλήρες σύνολο δεδομένων και έτσι μπορούν να υπολογιστούν χρησιμοποιώντας την εξίσωση (4.30). Επαναλαμβάνεται αυτή η επανεκτίμηση των τιμών για όλες τις μη παρατηρημένες τιμές στο D, παράγοντας ένα νέο πλήρες σύνολο δεδομένων D'_c. Υπολογίζεται η εκ των υστέρων κατανομής πιθανότητας $P(\theta^*_m|D'_c, M)$. Τέλος, επαναλαμβάνονται τα προηγούμενα τρία βήματα, και χρησιμοποιείται ο μέσος όρος του $P(\theta^*_M|D'_c, M)$ σαν προσέγγιση.

5.2 Laplace Προσέγγιση

Αυτή η μέθοδος ανήκει στην κατηγορία των μεθόδων μεγάλου δείγματος και είναι η πιο διαδεδομένη από όλες. Η βασική ιδέα πίσω από τις προσεγγίσεις μεγάλου δείγματος είναι ότι όσο το δείγμα N αυξάνει, ο όρος $P(\Theta_M|D, M) \propto P(D|\Theta_M, M) \cdot P(\Theta_M|M)$ μπορεί να προσεγγιστεί σαν μία πολυδιάστατη Gaussian κατανομή. Συγκεκριμένα, θέτουμε

$$g(\Theta_M) \equiv \log(P(D|\Theta_M, M) \cdot P(\Theta_M|M)). \tag{5.3}$$

Επίσης, ορίζεται $\widetilde{\Theta}_M$, να είναι η τιμή του Θ_M που μεγιστοποιεί την $g(\Theta_M)$. Αυτή η τιμή επίσης μεγιστοποιεί την $P(\Theta_M|D, M)$, και είναι γνωστή σαν *maximum a posteriori* (MAP) τιμή του Θ_M δοθέντος του D. Χρησιμοποιώντας πολυώνυμο Taylor δευτέρου βαθμού του $g(\Theta_M)$ γύρω από το $\widetilde{\Theta}_M$ για να προσεγγιστεί η $g(\Theta_M)$, προκύπτει

$$g(\Theta_m) \approx g(\widetilde{\Theta}_M) - \frac{1}{2}(\Theta_M - \widetilde{\Theta}_M)A(\Theta_M - \widetilde{\Theta}_M)^T, \tag{5.4}$$

όπου A ο αρνητικός Hessian της $g(\Theta_M)$ υπολογισμένος στο $\widetilde{\Theta}_M$. Υψώνοντας την $g(\Theta_M)$ στην εκθετική και χρησιμοποιώντας την εξίσωση (5.3) προκύπτει

$$P(D|\Theta_M, M)P(\Theta_M|M) \approx P(D|\widetilde{\Theta}_M, M)P(\widetilde{\Theta}_M|M)\exp\{-\frac{1}{2}(\Theta_M - \widetilde{\Theta}_M)A(\Theta_M - \widetilde{\Theta}_M)^T\}. \tag{5.5}$$

Έτσι, η προσέγγιση του $P(\Theta_M|D, M) \propto P(D|\Theta_M, M) \cdot P(\Theta_M|M)$ είναι Gaussian. Ολοκληρώνοντας και τις δύο πλευρές ως προς Θ_M και παίρνοντας τον λογάριθμο έχουμε την προσέγγιση

$$\log P(D|M) \approx \log P(D|\widetilde{\Theta}_M, M) + \log P(\widetilde{\Theta}_M|M) + \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|A|, \tag{5.6}$$

όπου d η διάσταση του M, δηλαδή ο αριθμός των παραμέτρων στο Θ_M. Αυτή η προσέγγιση για ολοκλήρωση είναι γνωστή σαν Laplace προσέγγιση. Οι Kass *et al.* [47] έχουν δείξει, ότι υπό ορισμένες προυποθέσεις, το σχετικό σφάλμα της προσέγγισης είναι $O_p(1/N)$, όπου N ο αριθμός των δειγμάτων στο D. Έτσι η προσσέγιση Laplace μπορεί να είναι πολύ ακριβής για μεγάλο N.

5.3 Προβλήματα των παραπάνω προσεγγίσεων

Εδώ κάνουμε μια σύνοψη των λόγων για τους οποίους οι δύο παραπάνω προσεγγίσεις δεν κάνουν για το πρόβλημα που αντιμετωπίζουμε. Ας θυμηθούμε τους περιορισμούς που

τίθονται από το πρόβλημα

1. Έχουμε μεγάλα δίκτυα γονιδίων.

2. Τα δεδομένα μας δεν είναι «επαρκή».

Από τα παραπάνω μπορούμε να καταλάβουμε γιατί και οι δύο προσεγγίσεις δεν ταιριάζουν στο πρόβλημα μας. Οι Monte-Carlo προσεγγίσεις έχουν μεγάλο υπολογιστικό κόστος και επομένως δεν μπορούν να χρησιμοποιηθούν για μεγάλα δίκτυα. Επίσης οι προσεγγίσεις μεγάλου δείγματος προϋποθέτουν την εύρεση των MAP παραμέτρων για τον υπολογισμό της πιθανοφάνειας κάθε υποψήφιου δικτύου άρα είναι και αυτές αρκετά χρονοβόρες. Πράγμα που φαίνεται και από το γεγονός ότι η απόδοση τους έχει εκτιμηθεί μόνο για μικρά δίκτυα. Σημαντικότερο ίσως είναι ότι τα δεδομένα μας όπως αναφέραμε δεν επαρκούν για να δώσουν αυτές οι προσεγγίσεις καλά αποτελέσματα. Καταφεύγουμε λοιπόν στην προσέγγιση του Friedman [48] που ονομάζεται Structural EM.

5.4 Structural EM (SEM)

Η γενική ιδέα είναι παρόμοια με αυτήν του κλασικού EM, δηλαδή γίνεται προσπάθεια αντί να μεγιστοποιηθεί η πιθανοφάνεια των δεδομένων να μεγιστοποιηθεί η αναμενόμενη πιθανοφάνεια. Η πρώτη προσέγγιση είναι η εύρεση της καλύτερης δομής κάνοντας την διαδικασία της αναζήτησης μέσα στην διαδικασία του EM. Έχοντας σαν δεδομένο ένα αρχικό δίκτυο υπολογίζονται οι MAP παράμετροί του, χρησιμοποιώντας EM ή gradient ascent και τα αναμενόμενα στατιστικά που χρειάζονται για να εκτιμήσουμε τα υποψήφια δίκτυα. Αφού αντικαθίστανται οι ελλιπείς τιμές με τα αναμενόμενα στατιστικά, η αναζήτηση γίνεται σε πλήρη δεδομένα και μπορούν να χρησιμοποιηθούν οι ιδιότητες των μέτρων αξιολόγησης για αποτελεσματική αναζήτηση.

Δηλώνονται ως **H** οι χαμένες τιμές και **O** οι παρατηρημένες. Όπως και πριν υποθέτεται ότι έχουμε μια κλάση μοντέλων που παραμετροποιείται από ένα διάνυσμα Θ_M, τέτοιο ώστε κάθε επιτρεπτή επιλογή τιμών για το Θ_M να δηλώνει μια κατανομή πιθανότητας στο σύνολο μεταβλητών $\mathcal{U}$. Επίσης έχουμε εκ των προτέρων πιθανότητα στα μοντέλα και στις παραμέτρους. Για ευκολία υποθέτουμε ότι έχουμε διακριτές μεταβλητές, ωστόσο τα αποτελέσματα εφαρμόζονται εύκολα και σε συνεχείς μεταβλητές με κάποιους περιορισμούς ομαλότητας και συνέχειας στις συναρτήσεις πιθανοφάνειας στα μοντέλα $\mathfrak{M}$. Όπως ήδη αναφέραμε δοθέντος ενός Bayesian δικτύου κάθε μεταβλητή εξαρτάται μόνο από τους πατέρες της στο γράφημα σύμφωνα με την υπόθεση 4.2. Εδώ κάθε παράγοντας f_i ανήκει στην εκθετική οικογένεια και μπορεί να γραφεί στην μορφή

$$f(\mathbf{X};\Theta) = e^{t(\Theta)\cdot s(\mathbf{X})} \tag{5.7}$$

όπου $t(\Theta)$ και $s(\mathbf{X})$ συναρτήσεις με διανυσματικές τιμές ίδιας διάστασης και $\cdot$ το εσωτερικό τους γινόμενο. Στην περίπτωση των πολυνυμικών παραγόντων μπορούν να γραφούν ως

εξής:

$$t(\Theta_{i,Pa_i}) = \langle \log \theta_{v_1,Pa_i}, ..., \log \theta_{v_l,Pa_i} \rangle, \quad (5.8)$$

$$s(\mathbf{x}) = \langle 1_{v_1,Pa_i}(\mathbf{x}), ..., 1_{v_l,Pa_i}(\mathbf{x}) \rangle, \quad (5.9)$$

όπου $v_1, ..., v_l$ οι πιθανές τιμές του X_i και $1_{\mathbf{y}}(\mathbf{x})$ είναι 1 αν οι τιμές των μεταβλητών $\mathbf{Y} \subseteq \mathbf{X}$ στο $\mathbf{y}$ ταιριάζουν με αυτές του $\mathbf{x}$, αλλιώς 0.

Πριν προχωρήσουμε θα κάνουμε ακόμα δύο υποθέσεις για το μοντέλο μας εκτός από τις (4.1) και (4.3).

Υπόθεση 5.1. Για κάθε μοντέλο $M \in \mathfrak{M}$ με k παράγοντες, η εκ των προτέρων κατανομή πάνω στις παραμέτρους έχει την μορφή

$$P(\Theta_1^M, \Theta_2^M, ...\Theta_k^M | M) = \prod_i^k P(\Theta_i^M | M). \quad (5.10)$$

Υπόθεση 5.2. Αν $f_i^M = f_i^{M'}$ για κάποια $M, M' \in \mathfrak{M}$, τότε

$$P(\Theta_i^M | M) = P(\Theta_i^{M'} | M'). \quad (5.11)$$

Υπό τις προϋποθέσεις 4.1, 4.3, 5.1, 5.2 και ένα σύνολο δεδομένων $D = \{u^1, ..., u^N\}$ πλήρων αναθέσεων τιμών στο $\mathcal{U}$, η εκ των υστέρων πιθανότητα των δεδομένων δοθέντος ενός μοντέλου M που περιέχει n παράγοντες-μεταβλητές $f_1, ..., f_n$ μπορεί να γραφεί ως εξής:

$$P(D|M) = \prod_{i=1}^{n} F_i(\sum_{j=1}^{N} s_i(u^j)), \quad (5.12)$$

όπου

$$F_i(S) = \int e^{t_i(\Theta_i) \cdot S} P(\Theta_i) d\Theta_i. \quad (5.13)$$

Τα $t_i(\cdot)$ και $s_i(\cdot)$ η εκθετική αναπαράσταση του παράγοντα f_i.

Για να βρούμε ένα MAP μοντέλο αρκεί να μεγιστοποιηθεί η πιθανότητα $P(D|M)P(M)$, αφού ο παράγοντας κανονικοποίησης είναι ίδιος για όλα τα μοντέλα που συγκρίνονται. Όπως ήδη αναφέραμε αν το σύνολο D των δεδομένων δεν είναι πλήρες δεν μπορεί να υπολογιστεί η πιθανότητα $P(D|M)$ αποτελεσματικά. Στην συνέχεια υποθέτεται ότι μπορεί να υπολογιστεί ή έστω να εκτιμηθεί η πιθανότητα του πλήρες συνόλου δεδομένων $P(\mathbf{H}, \mathbf{O}|M)$ (υπό τις προϋποθέσεις 4.1, 4.3, 5.1, 5.2). Επίσης γίνεται η υπόθεση ότι για ένα συγκεκριμένο μοντέλο μπορεί να υπολογιστεί αποτελεσματικά η πιθανότητα πρόβλεψης του επόμενου γεγονότος

$$P(X|M, D) = \int P(X|M, \Theta) P(\Theta|M, D) d\Theta \quad (5.14)$$

Αν και αυτό δεν ισχύει σε παραγοντικά μοντέλα, μπορεί να υπολογιστεί αποτελεσματικά μια προσέγγιση των παραπάνω προβλέψεων, χρησιμοποιώντας την MAP προσέγγιση.

Αλγόριθμος 5 Structural EM

Για $i = 1...\ell$ μέχρι την σύγκλιση

Υπολογίζεται η εκ των υστέρων πιθανότητα $P(\Theta_{M^{(i)}}|M, o)$

E-Step: Για κάθε M υπολογίζεται

$$Q(M : M^{(i)}) = E[\log P(\mathbf{H}, \mathbf{o}, M)|M^{(i)}, \mathbf{o}] = \sum_h P(\mathbf{h}|M^{(i)}, \mathbf{o}) \log P(\mathbf{h}, \mathbf{o}, M)$$

M-Step: Επιλέγεται το μοντέλο $M^{(i+1)}$ που μεγιστοποιεί $Q(M : M^{(i)})$

Αν $Q(M : M^{(i+1)}) = Q(M : M^{(i+1)})$

return $M^{(i)}$

Ο Αλγόριθμος

Έστω $P(H, O|M)$ η πιθανοφάνεια του πλήρους συνόλου δεδομένων, ο ψευδοκώδικας για τον αλγόριθμο είναι

5.4.1 Σύγκλιση του Structural EM

Το παρακάτω θεώρημα μας δείχνει πως κάθε επανάληψη δίνει και μια καλύτερη προσέγγιση.

Θεώρημα 5.1. *Αποδεικνύεται ότι αν $M^{(0)}, M^{(1)}, \ldots, M^{(i)}$ μια ακολουθία δικτύων που βρίσκει ο SEM τότε*

$$\log P(o, M^{(i+1)}) - \log P(o, M^{(i)}) \geq Q(M^{(i+1)} : M^{(i)}) - Q(M^{(i)} : M^{(i)}) \tag{5.15}$$

Απόδειξη

$$\begin{aligned}
\log \frac{P(\mathbf{h}, \mathbf{o}, M^{(i+1)})}{P(\mathbf{o}, M^{(i)})} &= \log \sum_{\mathbf{h}} \frac{P(\mathbf{h}, \mathbf{o}, M^{(i+1)})}{P(\mathbf{o}, M^{(i)})} \cdot \frac{P(\mathbf{h}|\mathbf{o}, M^{(i)})}{P(\mathbf{h}|\mathbf{o}, M^{(i)})} \\
&= \log \sum_{\mathbf{h}} P(\mathbf{h}|\mathbf{o}, M^{(i)}) \frac{P(\mathbf{h}, \mathbf{o}, M^{(i+1)})}{P(\mathbf{h}, \mathbf{o}, M^{(i)})} \\
&\geq \sum_{\mathbf{h}} P(\mathbf{h}|\mathbf{o}, M^{(i)}) \log \frac{P(\mathbf{h}, \mathbf{o}, M^{(i+1)})}{P(\mathbf{h}, \mathbf{o}, M^{(i)})} \\
&= E[\log \frac{P(\mathbf{H}, \mathbf{o}, M^{(i+1)})}{P(\mathbf{H}, \mathbf{o}, M^{(i)})}|M^{(i)}, \mathbf{o}] \\
&= Q(M^{(i+1)} : M^{(i)}) - Q(M^{(i)} : M^{(i)}).
\end{aligned}$$

Οπότε από το θεώρημα προκύπτει ότι αν σε κάθε βήμα βρίσκεται ένα δίκτυο που μεγιστοποιεί ή έστω έχει μεγαλύτερο αναμενόμενο μέτρο αξιολόγησης, τότε αποδεδειγμένα γίνεται μια καλύτερη επιλογή, με την έννοια του μέτρου αξιολόγησης περιθωρίου του δικτύου. Αυτό οδηγεί σε μια πιο χαλαρή εκδοχή του αλγορίθμου όπου αλλάζουμε το **M-step**:

Ορισμός 5.1. M-step Επιλέγεται ένα $M^{(i+1)}$ τέτοιο ώστε

$$Q(M^{(i+1)} : M^{(i)}) > Q(M^{(i)} : M^{(i)})$$

Το θεώρημα 5.2 δείχνει ότι η διαδικασία συγκλίνει όταν δεν υπάρχει επιπλέον βελτίωση στο μέτρο αξιολόγησης που στοχεύεται. Σαν έμμεση συνέπεια, μπορεί να δειχθεί ότι η διαδικασία φτάνει σε ένα τέτοιο σημείο υπό φυσιολογικές γενικά συνθήκες.

Θεώρημα 5.2. *Έστω $M^{(0)}, M^{(1)}, \ldots, M^{(i)}$ μια ακολουθία δικτύων που βρίσκει ο SEM. Αν ο αριθμός των μοντέλων στο είναι πεπερασμένος ή υπάρχει μια σταθερά c τέτοια ώστε $P(D|M, \Theta_M) < c$ για όλα τα μοντέλα M και τις παραμέτρους Θ_M, τότε υπάρχει το όριο $\lim_{n\to\infty} P(\mathbf{o}, M^{(i)})$.*

Δυστυχώς, δεν μπορεί να υπάρχει γνώση για το σημείο σύγκλισης. Για τον κλασικό EM, σημεία σύγκλισης είναι τα στατικά σημεία της συνάρτησης που θέλουμε να μεγιστοποιήσουμε. Δεν υπάρχει κάτι αντίστοιχο στον διακριτό χώρο των μοντέλων που γίνεται η αναζήτηση. Στην πραγματικότητα, το πιο προβληματικό σημείο του αλγορίθμου είναι ότι μπορεί να συγκλίνει σε κάποιο υπό-βέλτιστο μοντέλο. Αυτό μπορεί να συμβεί εάν το μοντέλο δημιουργεί μια κατανομή που καθιστά τα υποψήφια μοντέλα χειρότερα όταν εξετάζουμε το αναμενόμενο μέτρο αξιολόγησης. Διαισθητικά αναμένεται τέτοια φαινόμενα να είναι πιο συχνά καθώς ο λόγος της ελλιπής πληροφορίας προς τα δεδομένα αυξάνει. Μια πιθανή λύση είναι η επανάληψη του αλγορίθμου από πολλά διαφορετικά σημεία έναρξης.

Παραγοντική ιδιότητα

Μια πολύ σημαντική παρατήρηση που στην ουσία είναι και το μεγάλο πλεονέκτημα του SEM είναι η εφαρμογή του σε παραγοντικά μοντέλα. Αποδεικνύεται ότι

$$E[\log P(H, o, M) = \sum_{i=1}^{n} E[\log F_i(S_i)], \tag{5.16}$$

όπου F_i οποιαδήποτε εκθετική κατανομή και S_i οι τιμές των μεταβλητών της κατανομής. Το σημαντικό με αυτήν την ιδιότητα είναι ότι μπορεί να εφαρμοστεί οποιοσδήποτε ευρετικός αλγόριθμος με τον ίδιο τρόπο που χρησιμοποιείται στην περίπτωση με πλήρη δεδομένα. Βασιζόμενοι σε αυτήν, ο γενικός Αλγόριθμος 5 μπορεί να εξειδικευτεί για παραγοντικά μοντέλα δίνοντας τον Αλγόριθμο 6.

5.4.2 Προσεγγίσεις του $E[\log F_i(S_i)]$

Ένα σημαντικό ζήτημα που μένει υπό συζήτηση είναι πως υπολογίζεται το $E[\log F_i(S_i)]$. Η πιο απλή προσέγγιση είναι η

$$E[\log F_i(S_i)] = \log F_i(E[S_i]) \tag{5.17}$$

που ισχύει με την προϋπόθεση ότι η συνάρτηση $\log F_i(S_i)$ είναι γραμμική ως προς S_i. Ο Friedman [48] έδειξε ότι γενικότερα αποτελεί μια καλή προσέγγιση, αλλά μπορεί να προσεγγισθεί και με άλλους τρόπους. Το ερώτημα που τίθεται είναι πως υπολογίζονται οι

Αλγόριθμος 6 Structural EM - Παραγοντικά Μοντέλα

repeat

Για $i = 1...\ell$ μέχρι να φτάσουμε σε σημείο σύγκλισης.

Υπολογίζουμε τις MAP παραμέτρους $\hat{\Theta}^{M^{(i)}}$ για $M^{(i)}$ και O.

Πραγματοποιούμε αναζήτηση στα υποψήφια δίκτυα, αξιολογώντας το καθένα με

$$Score(M : M^{(i)}) = \sum_i E[\log F_i^M(S_i^M)|O, M^{(i)}, \hat{\Theta}^{M^{(i)}}]$$

Έστω $M^{(i+1)}$ το μοντέλο με το μεγαλύτερο μέτρο από αυτά που αξιολογήθηκαν.

until $Score(M^{(i)} : M^{(i)}) < Score(M^{(i+1)} : M^{(i)})$

return $M^{(i)}$.

αναμενόμενες τιμές των στατιστικών. Η προσέγγιση που χρησιμοποιείται είναι η εύρεση των MAP παραμέτρων βασιζόμενοι στο μοντέλο $M^{(i)}$ και η χρησιμοποίησή τους για τον υπολογισμό του αναμενόμενου μέτρου αξιολόγησης του μοντέλο $M^{(i+1)}$. Οι MAP παράμετροι δεν μπορούν να βρεθούν όπως είδαμε στην παράγραφο 4.2.1, αλλά χρησιμοποιούμε μεθόδους όπως οι gradient ascent ή ο EM, τις οποίες μεθόδους εξετάζουμε στις παραγράφους 5.5.1, 5.5.2.

5.5 Εκπαίδευση Παραμέτρων

Για να υπολογιστούν οι προσεγγίσεις που περιγράψαμε, πρέπει να καθοριστούν οι MAP ή ML τιμές του Θ_M. Αυτό μπορεί να επιτευχθεί με δύο τρόπους. Ο πρώτος είναι να χρησιμοποιήσουμε gradient ascent μεθόδους, όπου ακολουθείται η παράγωγος του $P(\Theta_M|D, M)$ ή $P(D|\Theta_M, M)$ σε ένα τοπικό ελάχιστο. Ο άλλος τρόπος είναι ο κλασικός EM αλγόριθμος. Θα παρουσιαστούν και οι δύο στην συνέχεια.

5.5.1 Gradient Ascent

Σε αυτήν την παράγραφο εξετάζεται πως μπορούμε να βρούμε τις παραγώγους των παραμέτρων ενός Bayesian δικτύου [49]. Καταρχάς η προσέγγιση αυτή στηρίζεται στο γεγονός ότι μπορούμε να δούμε την πιθανότητα $P_\theta(D)$ σαν μια συνάρτηση των παραμέτρων θ. Αυτό μειώνει το πρόβλημα στην εύρεση του μέγιστου μιας πολυδιάστατης μη γραμμικής συνάρτησης. Οι αλγόριθμοι για την επίλυση αυτού του προβλήματος κατά κανόνα ακολουθούν ένα μονοπάτι σε μια επιφάνεια όπου «συντεταγμένες» είναι οι παράμετροι , «ύψος» η τιμή της συνάρτησης και προσπαθούν να βρεθούν στο «ψηλότερο» σημείο στης επιφάνειας. Στην πραγματικότητα είναι ευκολότερο να μεγιστοποιηθεί η συνάρτηση πιθανοφάνειας $\log P_\theta(D)$. Από την στιγμή που οι συναρτήσεις είναι μονότονες, η μεγιστοποίηση της μίας συνεπάγεται και την μεγιστοποίηση της άλλης. Η πιο απλή μορφή αυτής της προσέγγισης είναι ο gradient ascent (hill-climbing). Σε κάθε σημείο θ, υπολογίζεται η παράγωγος $\nabla\theta$, το διάνυσμα των μερικών παραγώγων της συνάρτησης πιθανοφάνειας ως προς τις παραμέτρους. Ο

αλγόριθμος τότε κάνει ένα μικρό βήμα στην διεύθυνση της παραγώγου στο σημείο $\theta + \alpha\nabla\theta$, όπου α μία παράμετρος για το μέγεθος του βήματος. Αυτός ο αλγόριθμος συγκλίνει σε ένα τοπικό ελάχιστο για αρκετά μικρό α. Πρέπει να ληφθούν υπόψη οι περιορισμοί που τίθονται από την φύση των παραμέτρων που αναπαριστούν μια συνάρτηση πιθανότητας. Αυτοί είναι ότι $\theta_{ijk} \in [0,1]$ και $\sum_{j=1}^{r_i} \theta_{ijk} = 1$. Αυτοί οι περιορισμοί μπορούν να ικανοποιηθούν προβάλλοντας το $\nabla\theta$ σε μια περιορισμένη επιφάνεια. Γενικά, η προβολή επιτυγχάνεται παίρνοντας το κανονικοποιημένο διάνυσμα στην επιφάνεια περιορισμών, προβάλλοντας το διάνυσμα παραγώγων σε αυτό και αφαιρώντας το αρχικό διάνυσμα από αυτό. Στην περίπτωση αυτή, η επιφάνεια περιορισμών απαιτεί $\sum_{j=1}^{r_i} \theta_{ijk} = 1, \quad \forall i,k$. Αν για κάποια τιμή των i και k έχουμε J διαφορετικές τιμές για το j, το διάνυσμα που είναι ορθοκανονικό στην επιφάνεια περιορισμών έχει τιμή $1/\sqrt{J}$ για κάθε ένα από τα σχετικά στοιχεία. Έτσι, η προβολή της παραγώγου θα έχει στις θέσεις που αντιστοιχούν στο ijk, το μέσο των στοιχείων της παραγώγου που αντιστοιχούν στο $\theta_{i1k}, ..., \theta_{iJk}$. Αφαιρώντας αυτό το διάνυσμα από το αρχικό διάνυσμα παραγώγου προκύπτει το κανονικοποιημένο διάνυσμα. Παρατηρούμε ότι στο κανονικοποιήμενο διάνυσμα παραγώγου, το άθροισμα που αντιστοιχεί στα στοιχεία θ_{ijk} για συγκεκριμένα i,j και k είναι μηδέν. Έτσι αν κάνουμε ένα μικρό βήμα κατά μήκος του διανύσματος, το άθροισμα $\sum_{j=1}^{r_i} \theta_{ijk}$ θα παραμείνει αμετάβλητο. Οπότε, αν ξεκινήσουμε από ένα σημείο της επιφάνειας περιορισμών, απαραίτητα παραμένουμε σε αυτήν, όπως είναι επιθυμητό. Μια εναλλακτική μέθοδος για την ικανοποίηση των περιορισμών που χρησιμοποιείται ευρέως στην στατιστική, είναι να εισαχθούν κάποιες υπερπαραμέτροι που από κατασκευής ικανοποιούν τους περιορισμούς. Σε αυτήν την περίπτωση εισάγονται οι παράμετροι w_{ijk}

$$w_{ijk} = \frac{\theta_{ijk}^2}{\sum_{j'=1}^{r_i} \theta_{ij'k}^2} \tag{5.18}$$

Από τον τύπο φαίνεται εύκολα πως ικανοποιούνται οι περιορισμοί μας. Επιπλέον ένα τοπικό μέγιστο ως προς το w_{ijk} είναι επίσης τοπικό μέγιστο και για το θ_{ijk} και αντίστροφα. Οπότε μπορούμε να βελτιστοποιήσουμε ως προς το θ_{ijk}. Η παράγωγος μπορεί να βρεθεί υπολογίζοντας τη (μη κανονικοποιημένη) παράγωγο ως προς το θ_{ijk} και στην συνέχεια βρίσκοντας την παράγωγο ως προς το w_{ijk} με τον κανόνα της αλυσίδας.

Τοπικοί υπολογισμοί της παραγώγου

Η χρησιμότητα των μεθόδων βασισμένων σε παραγώγους εξαρτάται από την δυνατότητα να υπολογίσουμε την παράγωγο αποτελεσματικά. Αυτός είναι και ο λόγος της ευρείας χρήσης του gradient descent στα νευρωνικά δίκτυα. Παρακάτω δείχνουμε μια παρόμοια κατάσταση που ισχύει και για πιθανοτικά δίκτυα. Στην πραγματικότητα, για πιθανοτικά δίκτυα, ο αλγόριθμος συμπερασματολογίας (inference), πραγματοποιεί όλους τους απαραίτητους υπολογισμούς. Η παράγωγος μπορεί να υπολογισθεί τοπικά για κάθε κόμβο χρησιμοποιώντας πληροφορία από τον αλγόριθμο συμπερασματολογίας (Κεφάλαιο 6).

Υπολογισμός της παραγώγου

Παρακάτω δείχνεται πως μπορεί να υπολογιστεί η συνεισφορά κάθε περίπτωσης στην παράγωγο ξεχωριστά. Στην συνέχεια αθροίζονται τα αποτελέσματα.

$$\begin{aligned}\frac{\partial \ln P_\theta(D)}{\partial \theta_{ijk}} &= \frac{\partial \ln \prod_{l=1}^{m} P_\theta(D_l)}{\partial \theta_{ijk}} \\ &= \sum_{l=1}^{m} \frac{\partial \ln P_\theta(D_l)}{\partial \theta_{ijk}} \\ &= \sum_{l=1}^{m} \frac{\partial \ln P_\theta(D_l)/\partial \theta_{ijk}}{P_\theta(D_l)}.\end{aligned} \tag{5.19}$$

Τώρα στόχος είναι να βρεθεί ένα απλός αλγόριθμος για να υπολογιστεί κάθε μία από τις εκφράσεις $\frac{\partial \ln P_\theta(D_l)/\partial \theta_{ijk}}{P_\theta(D_l)}$. Για να πάρουμε μία έκφραση σε όρους της τοπικής πληροφορίας για τις παραμέτρους θ_{ijk}, εισάγονται τα X_i και $\mathbf{Pa}_i$ παίρνοντας τον μέσο όρο των πιθανών τιμών τους:

$$\begin{aligned}\frac{\partial \ln P_\theta(D_l)/\partial \theta_{ijk}}{P_\theta(D_l)} &= \frac{\frac{\partial}{\partial \theta_{ijk}}(\sum_{j',k'} P_\theta(D_l|x_{ij'}, pa_{ik'})P_\theta(x_{ij'}, pa_{ik'}))}{P_\theta(D_l)} \\ &= \frac{\frac{\partial}{\partial \theta_{ijk}}(\sum_{j',k'} P_\theta(D_l|x_{ij'}, pa_{ik'})P_\theta(x_{ij'}|pa_{ik'})P_\theta(pa_{ik'}))}{P_\theta(D_l)}.\end{aligned} \tag{5.20}$$

Η βασική ιδιότητα αυτής της έκφρασης είναι ότι ο όρος θ_{ijk} εμφανίζεται μόνο σε γραμμική μορφή. Στην πραγματικότητα, ο όρος θ_{ijk} εμφανίζεται μόνο σε έναν όρο στο άθροισμα, για $j' = j$, $k' = k$. Για αυτόν το όρο, $P_\theta(x_{ij'}|pa_{ik'})$ είναι απλά θ_{ijk}. Έτσι

$$\begin{aligned}\frac{\partial P_\theta(D_l)/\partial \theta_{ijk}}{P_\theta(D_l)} &= \frac{P_\theta(D_l|x_{ij}, pa_{ik})P_\theta(pa_{ik})}{P_\theta(D_l)} \\ &= \frac{P_\theta(x_{ij}, pa_{ik}|D_l)P_\theta(D_l)P_\theta(pa_{ik})}{P_\theta(x_{ij}, pa_{ik})P_\theta(D_l)} \\ &= \frac{P_\theta(x_{ij}, pa_{ik}|D_l)}{P_\theta(x_{ij}|pa_{ik})} \\ &= \frac{P_\theta(x_{ij}, pa_{ik}|D_l)}{\theta_{ijk}}.\end{aligned} \tag{5.21}$$

Κάθε αλγόριθμος συμπερασματολογίας για πιθανοτικά δίκτυα, όταν εκτελείται με δεδομένα D_l, υπολογίζει και τον όρο $P_\theta(x_{ij}, pa_{ik}|D_l)$. Για παράδειγμα ο αλγόριθμος Συνδετικών δέντρων (Junction tree) των Lauritzen και Spiegelhalter [50], υπολογίζει την εκ των υστέρων πιθανότητα για κάθε κλίκα στο συνδετικό δέντρο που αντιστοιχεί στο αρχικό Bayesian δίκτυο. Από την στιγμή που κάθε κόμβος και οι πατέρες του εμφανίζονται τουλάχιστον σε μία κλίκα, οι αντίστοιχες πιθανότητες μπορούν να βρεθούν απαλείφοντας όλες τις άλλες μεταβλητές στην κλίκα.

5.5.2 Υπολογισμός παραμέτρων με EM

Στην αρχή τίθεται στο σύνολο παραμέτρων Θ_M μία τυχαία αρχική τιμή. Στην συνέχεια υπολογίζονται τα αναμενόμενα επαρκή στατιστικά για ένα πλήρες σύνολο δεδομένων, όπου η αναμενόμενη τιμή υπολογίζεται ως προς την από κοινού κατανομή του συνόλου $\mathbf{X}$ υπό την δοσμένη τιμή του Θ_M και τα γνωστά δεδομένα D. Για Bayesian δίκτυα με διακριτές μεταβλητές, υπολογίζουμε

$$E_{P(x|D,\Theta_M,M)}(N_{ijk}) = \sum_{l=1}^{N} P(x_{ik}, pa_{ij}|x_l, \Theta_M, M), \tag{5.22}$$

όπου x_l, είναι η πιθανή μη πλήρης l-οστή περίπτωση στο D. Όταν το X_i και όλες οι μεταβλητές στο σύνολο των πατέρων Pa_i είναι παρατηρημένες στην περίπτωση x_l, ο αντίστοιχος όρος είναι απλά μηδέν ή ένα. Σε άλλη περίπτωση μπορεί να χρησιμοποιηθεί ένας αλγόριθμος συμπερασματολογίας (inference) σε Bayesian δίκτυα. Αυτός ο υπολογισμός αποτελεί και το Expectation βήμα του EM αλγορίθμου. Στην συνέχεια χρησιμοποιούνται τα αναμενόμενα επαρκή στατιστικά, σαν να είναι τα πραγματικά επαρκή στατιστικά από ένα πλήρες σύνολο δειγμάτων D_c. Αν κάνουμε υπολογισμό των ML παραμέτρων, μπορούμε να καθορίσουμε τις τιμές των παραμέτρων Θ_M που μεγιστοποιούν $P(D_c|\Theta_M, M)$. Αυτές οι τιμές δίνονται από την σχέση

$$\theta_{ijk} = \frac{E_{P(x|D,\theta_m,m)}(N_{ijk})}{\sum_{k=1}^{r_i} E_{P(x|D,\theta_m,m)}(N_{ijk})}. \tag{5.23}$$

Αν θέλουμε τις MAP παραμέτρους, μπορούμε να καθορίσουμε τις τιμές του θ_m που μεγιστοποιούν την εκ των υστέρων πιθανότητα των παραμέτρων:

$$\theta_{ijk} = \frac{a_{ijk} + E_{P(x|D,\theta_m,m)}(N_{ijk})}{\sum_{k=1}^{r_i}(\alpha_{ijk} + E_{P(x|D,\theta_m,m)}(N_{ijk}))}. \tag{5.24}$$

Αυτό το βήμα ονομάζεται maximization βήμα του EM αλγορίθμου. Οι Dempster *et al.* [51] έδειξαν ότι η επανάληψη των δύο βημάτων συγκλίνει σε τοπικό μέγιστο. Ο EM αλγόριθμος μπορεί να εφαρμοστεί όταν υπάρχουν τα επαρκή στατιστικά (οι τοπικές συναρτήσεις κατανομής πιθανότητας ανήκουν στην εκθετική οικογένεια), παρόλο που υπάρχουν και επεκτάσεις του EM που εφαρμόζονται και σε πιο πολύπλοκες τοπικές κατανομές.

5.5.3 Μάθηση παραμέτρων με ελλιπή Δεδομένα

Σε αντιστοιχία με τις παραγράφους (4.2.1) και (4.2.2) που είχαμε πλήρη δεδομένα, παρακάτω περιγράφουμε πως εκπαιδεύουμε τις παραμέτρους για πολυνυμικές και γραμμικές Gaussian κατανομές όταν τα δεδομένα είναι ελλιπείς. Στην ουσία περιγράφεται το M-STEP του EM αλγορίθμου που είναι η ανανέωση των παραμέτρων, ενώ όπως είδαμε το E-STEP είναι η εύρεση των αναμενόμενων επαρκών στατιστικών (expected sufficient statistics). Στην περίπτωση που τα δεδομένα δεν είναι πλήρη αντικαθιστούμε τα επαρκή στατιστικά με τα

αναμενόμενα επαρκή στατιστικά. Στην διακριτή περίπτωση έχουμε

$$E[N_{ijk}] = \sum_{\ell} P(X_i = k, Pa_i = j|e_\ell), \tag{5.25}$$
$$= \sum_{\ell} \frac{P(X_i = k, Pa_i = j, e_\ell)}{P(e_\ell)}.$$

Η ανανέωση των τιμών των παραμέτρων για την ML περίπτωση δίνεται από την σχέση (5.23) και στην MAP περίπτωση από την σχέση (5.24). Στην συνεχή περίπτωση τα αναμενόμενα επαρκή στατιστικά είναι

$$s'_N = \sum_{\ell} E[X_\ell|e_\ell], \tag{5.26}$$

$$Q'_N = \sum_{\ell} E[X_\ell X_\ell^T|e_\ell]. \tag{5.27}$$

Απλά στις σχέσεις (4.14), (4.15) αντικαθιστούμε τα s_N, Q_N με τα s'_N και Q'_N αντίστοιχα.

Όπως ήδη αναφέραμε τόσο για την εκτίμηση των ML παραμέτρων όσο και για την εκτίμηση νέων υποψηφίων χρειαζόνται τα αναμενόμενα στατιστικά κάθε κατανομής του δικτύου. Για να υπολογιστούν πρέπει να εκτιμηθεί η πιθανότητα $P(x_{il}, pa_{il}|e_l)$ [52]. Για να εκτιμηθεί αυτή η πιθανότητα χρειάζεται ένα εργαλείο συμπερασματολογίας για Bayesian δίκτυα.

Κεφαλαιο 6

Συμπερασματολογια (Inference) σε Bayesian Δικτυα

Σαν συμπερασματολογία στα Bayesian δίκτυα ορίζεται ο υπολογισμός της πιθανότητας $P(X_Q|X_E = x_e)$ όπου X_Q ένα σύνολο μεταβλητών υπό ερώτηση και X_E οι γνωστές μεταβλητές . Για παράδειγμα, στην ιατρική διάγνωση, οι μεταβλητές X_E θα αντιπροσώπευαν τα παρατηρημένα συμπτώματα και οι μεταβλητές X_Q τα πιθανά αίτια αυτών των συμπτωμάτων. Αυτή η ποσότητα μπορεί να υπολογισθεί από την από κοινού κατανομή $P(X_1, ...X_N)$ χρησιμοποιώντας τον κανόνα του Bayes:

$$P(X_Q|X_E) = \frac{P(X_Q, X_E)}{P(X_E)} = \frac{\sum_{h \notin Q \cup E} P(X_H = h, X_Q, X_E)}{\sum_{h \notin E} P(X_H = h, X_E)}$$

'Ετσι στην ουσία η συμπερασματολογία ισοδυναμεί με τον υπολογισμό των περιθωριοποίησεων (marginals) των από κοινού κατανομών. Αν όλες οι μεταβλητές είναι δυαδικές, τότε ο υπολογισμός του $\sum_h P(X_1, ..., X_N)$ απαιτεί χρόνο $O(2^N)$. Η έρευνα επικεντρώνεται στο να γίνει ο παραπάνω υπολογισμός όσο το δυνατό πιο αποδοτικά. Αν η από κοινού κατανομή αναπαριστάται από ένα Bayesian δίκτυο, μπορεί να γραφεί σε παραγοντική μορφή ως εξής:

$$P(X_1, ..., X_N) = \prod_{i=1}^{N} P(X_i|Pa_i). \tag{6.1}$$

Τότε η περιθωριοποίηση μπορεί να γίνει πιο αποδοτικά «σπρώχνοντας τα αθροίσματα μέσα από τους παράγοντες». Για παράδειγμα θεωρούμε το Bayesian δίκτυο του Σχήματος 6.1.

Η από κοινού κατανομή πιθανότητας μπορεί να γραφεί ως εξής:

$$P(A,B,C,D,F,G) = P(A) * P(B|A) * P(C|A) * P(D|B,A) * P(F|B,C) * P(G|F)$$

Ας υποθέσουμε ότι θέλουμε να περιθωριοποιήσουμε όλες τις μεταβλητές. Προφανώς το αποτέλεσμα θα είναι 1, ωστόσο η ιδέα είναι ίδια και όταν θέλουμε να περιθωριοποιήσουμε ένα υποσύνολο των μεταβλητών. Αυτό μπορεί να γραφεί ως εξής:

$$\sum_{A,B,C,D,F,G} P(A,B,C,D,F,G) = \sum_A P(A) * \sum_B P(B|A) * \sum_C P(C|A) * \sum_D P(D|B,A) * \sum_F P(F|B,C) * \sum_G P(G|F)$$

Δουλεύοντας από δεξιά προς τα αριστερά, προκύπτει

$$\sum_A P(A) * P \sum_B (B|A) * \sum_C P(C|A) * \sum_D P(D|B,A) * \sum_F P(F|B,C) * \lambda_{G\to F}(F),$$

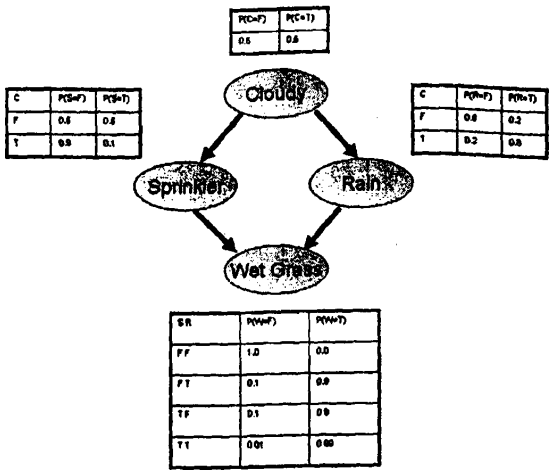
όπου $\lambda_{G\to F}(F) = \sum_G P(G|F)$. Προφανώς σε αυτήν την περίπτωση $\lambda_{G\to F}(F) = 1$ για κάθε F, αλλά αυτό μπορεί να μην ισχύει γενικά. Στο επόμενο βήμα παίρνουμε

$$\sum_A P(A) * P \sum_B (B|A) * \sum_C P(C|A) * \lambda_{F\to C}(B,C) * \sum_D P(D|B,A),$$

όπου $\lambda_{F\to C}(B,C) = \sum_F P(F|B,C) * \lambda_{G\to F}(F)$. Σημειώστε πως αυτός ο όρος μπαίνει μετά το άθροισμα πάνω στο C, παρακάμπτοντας το άθροισμα πάνω στο D. Γενικά οι όροι μετακινούνται όσο πιο αριστερά γίνεται, ώστε να ελαχιστοποιηθεί το υπολογιστικό κόστος. Ο περιορισμός έγκειται στο γεγονός ότι όλες οι μεταβλητές που εμφανίζονται σε έναν όρο πρέπει να είναι στην εμβέλεια του κατάλληλου τελεστή αθροίσματος. Με τον συμβολισμό $\lambda_{F\to C}(B,C)$ δηλώνεται ο όρος που προέρχεται από το άθροισμα πάνω στο F και πηγαίνει στο άθροισμα στο C. Πηγαίνει στο C και όχι στο B επειδή το C είναι ψηλότερα στην σειρά διαγραφής (άθροισης). Πρέπει να είναι σαφές ότι η σειρά με την οποία πραγματοποιούνται οι αθροίσεις, καθορίζει το μέγεθος των ενδιάμεσων όρων λ, το οποίο επηρεάζει δραστικά το υπολογιστικό φόρτο. Στην συνέχεια θα αποκαλούμε αυτήν την σειρά διαγραφής με τον όρο π. Μπορούμε να συνεχίσουμε με τον ίδιο τρόπο μέχρι να υπολογιστεί η επιθυμητή περιθωριοποίηση. Η ιδέα της κατανομής των αθροισμάτων στα γινόμενα έχει ανεξάρτητα μελετηθεί πολλές φορές και έχει πολλά ονόματα: peeling, symbolic probabilistic inference (SPI), variable elimination, bucket elimination κ.τ.λ.

6.1 Παράδειγμα Συπερασματολογίας

Έστω το δίκτυο του Σχήματος 6.1 και οι πίνακες κατανομής πιθανότητας που φαίνονται. Έστω ότι έχουμε παρατηρήσει ότι το γρασίδι είναι βρεγμένο.



Σχήμα 6.1: Παράδειγμα Δικτύου και των πινάκων κατανομής πιθανότητας

Υπάρχουν δύο πιθανές εξηγήσεις γι' αυτό. Να έχει βρέξει ή να έχει υγρασία. Θέλουμε δηλαδή να υπολογίσουμε τις πιθανότητες $P(S = 1|W = 1)$ και $P(R = 1|W = 1)$. Η από κοινού κατανομή με τον κανόνα της αλυσίδας είναι

$$P(C, S, R, W) = P(C) * P(S|C) * P(R|C, S) * P(W|C, S, R).$$

Εκμεταλλευόμενοι τις εξαρτήσεις από το δίκτυο έχουμε

$$P(C, S, R, W) = P(C) * P(S|C) * P(R|C) * P(W|S, P).$$

Τότε

$$\begin{aligned} P(S = 1|W = 1) &= \frac{P(S = 1, W = 1)}{P(W = 1)}, \\ &= \frac{\sum_{r,c} P(C = s, R = r, W = 1, S = 1)}{P(W = 1)}, \end{aligned}$$

$$\begin{aligned} P(R = 1|W = 1) &= \frac{P(S = 1, W = 1)}{P(W = 1)}, \\ &= \frac{\sum_{s,c} P(C = s, S = s, W = 1, R = 1)}{P(W = 1)}, \end{aligned}$$

όπου

$$P(W = 1) = \sum_{r,c,s} P(C = s, S = s, R = r, W = 1).$$

Για διακριτές μεταβλητές η από κοινού κατανομή έχει μέγεθος 2^n. Άρα προσθέτοντας κατευθείαν στην από κοινού κατανομή πιθανότητας έχει εκθετική πολυπλοκότητα. Γι'αυτό πρέπει να βρεθούν πιο αποδοτικοί τρόποι.

6.2 Συνδετικά Δέντρα

Ας υποθέσουμε ότι θέλουμε να υπολογίσουμε την πιθανότητα $P(X_i|X_E)$ για κάθε $i \notin E$. Θα μπορούσαμε να καλέσουμε την διαδικασία της διαγραφής των μεταβλητών (variable elimination) $O(N)$ φορές, μία για κάθε i, αλλά αυτό θα ήταν αναποτελεσματικό χωρίς λόγο, από την στιγμή που πραγματοποιούμαι πολλές φορές τις ίδιες πράξεις. Πιο συγκεκριμένα από την στιγμή που η διαγραφή απαιτεί χρόνο $O(N)$, N κλήσεις απαιτούν $O(N^2)$ χρόνο. Στην συνέχεια εξετάζεται ένα τρόπος υπολογισμού όλων των N περιθωριοποίησεων σε χρόνο $O(N)$. Η βασική ιδέα είναι η αποθήκευση των ενδιάμεσων όρων λ που δημιουργούνται καθώς δουλεύουμε από δεξιά προς τα αριστερά μέσα από τις αθροίσεις, και μετά η επαναχρησιμοποίηση τους καθώς δουλεύουμε από αριστερά προς τα δεξιά. Οι λ θα αποθηκευτούν σε μια δευτερεύουσα δομή δέντρου που ονομάζεται *Συνδετικό δέντρο* (Junction Tree).

Κατασκευή Συνδετικού δέντρου

Τα βασικά βήματα για την κατασκευή ενός Συνδετικού δέντρου από ένα αρχικό Bayesian δίκτυο είναι τα εξής:

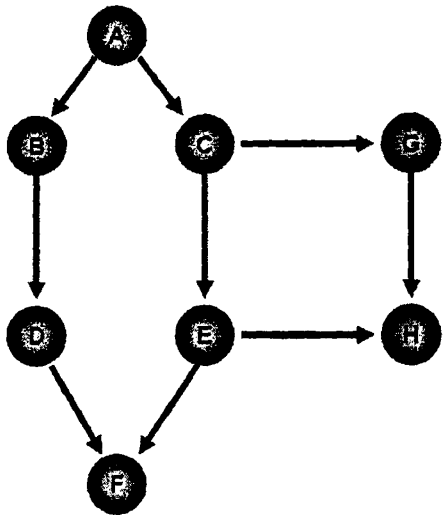
- Μη κατευθυνόμενο γράφημα (Moral Graph). Αναιρούνται οι κατευθύνσεις του γραφήματος. Ενώνονται οι πατέρες κάθε κόμβου.
- Τριγωνοποίηση (Triangulation). Δημιουργείται το τριγωνοποιημένο γράφημα, τέτοιο ώστε να μην υπάρχει κύκλος μήκους μεγαλύτερο του 3 .
- Maximum Cardinality Algorithm (Kruskal) ή κάποιος άλλος αλγόριθμος, για να βρεθεί η σειρά διαγραφής (elimination).
- Βρίσκονται οι κλίκες από την σειρά διαγραφής.
- Δημιουργία του Συνδετικού δέντρου από τις κλίκες.

6.2.1 Μη κατευθυνόμενο γράφημα (Moral Graph)

Αναιρούνται οι κατευθύνσεις στο αρχικό μας κατευθυνόμενο γράφημα και στην συνέχεια ενώνονται οι πατέρες κάθε κόμβου μεταξύ τους ανά δύο (για όσους δεν υπάρχει ήδη σύνδεση). Το αποτέλεσμα είναι ένα μη κατευθυνόμενο γράφημα. Αυτό το βήμα εξασφαλίζει ότι στο τελικό δέντρο για κάθε κόμβο θα υπάρχει τουλάχιστον μία κλίκα που θα περιέχει τον κόμβο και όλους τους πατέρες του.

6.2.2 Τριγωνοποίηση του γραφήματος

Καταρχάς με τον όρο τριγωνοποιημένο γράφημα δηλώνεται κάθε μη κατευθυνόμενο γράφημα αν για κάθε κύκλο μήκους μεγαλύτερο του 3 υπάρχει μια χορδή. Στόχος είναι από το moral γράφημα να δημιουργηθεί ένα τριγωνοποιημένο γράφημα, προσθέτοντας όσο το

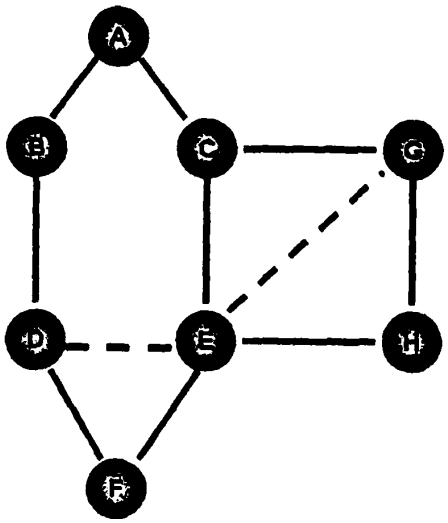


Σχήμα 6.2: Αρχικό Γράφημα

λιγότερο επιπρόσθετες ακμές. Η βιβλιογραφία πάνω στο θέμα είναι εκτενέστατη [53]. Η συνήθης τεχνική είναι η εύρεση μια σειράς (elimination, removing, triangulation) που να περιέχει όλους τους κόμβους στο γράφημα. Η μέθοδος αποτελείται από την διαδικασία διαγραφής (elimination), που ακολουθεί την σειρά επιλογής και αφαιρεί όλους τους κόμβους από το γράφημα. Αν με $adj(X_i)$ δηλώνεται το σύνολο των γειτόνων του X_i, στο μη κατευθυνόμενο γράφημα, τότε με τον όρο αφαίρεση του X_i δηλώνεται η διαδικασία με την οποία προσθέτοντας κατάλληλες ακμές μετατρέπουμε το υπογράφημα $X_i \bigcup adj(X_i)$ σε ένα πλήρες υπογράφημα και στην συνέχεια αφαιρείται μαζί με τις εισερχόμενες σε αυτό ακμές από το γράφημα. Το τριγωνοποιημένο γράφημα $\mathcal{G}_T$ θα είναι το αποτέλεσμα της πρόσθεσης των ακμών που εισάχθηκαν στην φάση της διαγραφής στο μη κατευθυνόμενο γράφημα. Στην συνέχεια πρέπει να βρεθούν οι κλίκες του τριγωνοποιημένου γραφήματος. Η διαδικασία που ακολουθείται είναι η εξής:

1. $\mathfrak{L} = \emptyset$
2. Αφαιρείται ένας κόμβος X_i (συνήθως σύμφωνα με την σειρά διαγραφής) από το γράφημα. Προστίθεται το $C_i = X_i \bigcup adj(X_i)$ στο $\mathfrak{L}$
3. Αν η C_i δεν περιέχει όλους τους εναπομείναντες κόμβους επαναλαμβάνεται το προηγούμενο βήμα.
4. «Πριονίζεται» το $\mathfrak{L}$ αφαιρώντας τα μη maximal σύνολα

Το παραγόμενο $\mathfrak{L}$ είναι το σύνολο των κλικών στο $\mathcal{G}_T$.



Σχήμα 6.3: Moral Γράφημα

6.2.3 Κατασκεύη του Συνδετικού δέντρου

Πριν περιγράψουμε την κατασκευή από το σύνολο των κλικών του συνδετικού δέντρου παραθέτουμε τον ορισμό του Συνδετικού δέντρου.

Ορισμός 6.1. Ένα *Συνδετικό Δέντρο* είναι ένα γράφημα με μορφή δέντρου στο οποίο οι κόμβοι του είναι τα στοιχεία του $\mathfrak{L}$, όπου $\mathfrak{L}$ το σύνολο των κλικών από ένα μη κατευθυνόμενο γράφημα, και στο οποίο ισχύει η ακόλουθη ιδιότητα:

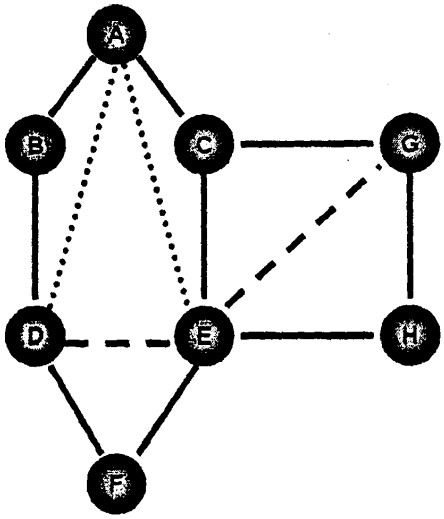
Για κάθε ζευγάρι κόμβων V και W στο δέντρο, όλοι οι κόμβοι στο μονοπάτι μεταξύ των V και W περιέχουν τις μεταβλητές που ανήκουν στο $V \cap W$.

Υπάρχει ένα συστηματικός τρόπος για την κατασκευή του συνδετικού δέντρου από το τριγωνοποιημένο γράφημα και το σύνολο των κλικών του.

6.2.4 Μέθοδοι Τριγωνοποίησης του γραφήματος

Όπως αναφέραμε το βασικότερο σημείο στον παραπάνω αλγόριθμο για την κατασκευή του JT είναι η επιλογή της σειράς διαγραφής (elimination) ώστε να πάρουμε όσο το λιγότερο επιπρόσθετες ακμές γίνεται. Παρόλο που μπορεί να επιλεγεί οποιαδήποτε σειρά από τις συνολικά $|\mathbf{X}|!$, το πρόβλημα εύρεσης της βέλτιστης σειράς αποδεικνύεται ότι είναι NP-hard πρόβλημα. Για αυτόν ακριβώς τον λόγο πολλοί ευρετικοί αλγόριθμοι έχουν προταθεί στην βιβλιογραφία. Χωρίζονται σε δύο βασικές κατηγορίες:

1. *Greedy μέθοδοι.* Αυτοί είναι ευρετικοί αλγόριθμοι που σε κάθε βήμα επιλέγεται η επόμενη μεταβλητή για να αφαιρεθεί σύμφωνα με κάποιο κριτήριο. Τα τρία πιο γνωστά κριτήρια είναι ο αριθμός των ακμών που προστίθενται όταν αφαιρείται η μεταβλητή X_i



Σχήμα 6.4: Τριγωνοποιημένο Γράφημα

΄(min-size), ο αριθμός των μεταβλητών που περιέχονται στην κλίκα που δημιουργείται από την αφαίρεση της μεταβλητής X_i (min-size), και το βάρος/μέγεθος της κλίκας που δημιουργείται από την αφαίρεση της μεταβλητής X_i (min-weight). Αυτοί οι αλγόριθμοι δίνουν καλές λύσεις σχετικά γρήγορα.

2. *Γενετικοί αλγόριθμοι και τεχνικές συνδυαστικής βελτιστοποίησης (combinatorial optimization).* Η ιδέα είναι να εφαρμοστεί ένας αλγόριθμος μεταευρετικής φύσης παίρνοντας σαν συνάρτηση κόστους την ελαχιστοποίηση του μεγέθους του χώρου καταστάσεων του συνδετικού δέντρου. Επομένως, κάθε πιθανή λύση αναπαριστάται σαν μια σειρά αφαίρεσης και το μέτρο αξιολόγησης της υπολογίζεται σαν το άθροισμα του μεγέθους των κλικών που παράγονται από την τριγωνοποίηση του γραφήματος χρησιμοποιώντας αυτήν την σειρά.

6.2.5 Maximum Cardinality Search (MCS)

Στην πρώτη κατηγορία αλγορίθμων ανήκει και ο MCS που χρησιμοποιήσαμε εμείς για την τριγωνοποίηση του γραφήματος [54]. Είναι ένας απλός αλγόριθμος γραμμικής πολυπλοκότητας που παίρνει σαν είσοδο ένα μη κατευθυνόμενο γράφημα και μας δίνει την αντίστροφη σειρά διαγραφής. Το πλεονέκτημα του είναι ότι είναι απλός και γρήγορος και δίνει σχετικά καλά αποτελέσματα.

Αλγόριθμος 7 MCS με είσοδο ένα γράφημα G και έξοδο την σειρά απαλοιφής π

1: BEGIN
2: Για όλους τους κόμβους u του γραφήματος G, $w(v) = 0$
3: For $i = n$ to 1 do
4: Επιλέγεται μια ακμή z που δεν έχει αριθμηθεί με το μεγαλύτερο βάρος. Θέτουμε $\alpha(z) = i$.
5: Για κάθε μη αριθμημένη μεταβλητή $y \in N(z)$ τίθεται $w(y) = w(y) + 1$
6: END

6.2.6 Δημιουργία Συνδετικού δέντρου από τις κλίκες

Από το σημείο αυτό δεν χρειάζεται το μη κατευθυνόμενο γράφημα. Ψάχνουμε να βρούμε το βέλτιστο δέντρο από τις κλίκες που έχουμε βρει. Για να δημιουργηθεί το βέλτιστο δέντρο, πρέπει να συνδεθούν οι κλίκες έτσι ώστε το παραγόμενο δέντρο των κλικών να ικανοποιεί την ιδιότητα του Συνδετικού δέντρου σύμφωνα με τον Ορισμό 6.1 και ένα κριτήριο βελτιστότητας που θα οριστεί παρακάτω. Το κριτήριο βελτιστότητας ευνοεί τα δέντρα που ελαχιστοποιούν το υπολογιστικό χρόνο που απαιτείται για την διαδικασία της συμπερασματολογίας. Δοθέντος ενός συνόλου κλικών, μπορεί να δημιουργηθεί ένα δέντρο κλικών εισάγοντας αναδρομικά ακμές μεταξύ των κλικών μέχρι οι κλίκες να είναι ενωμένες με $n - 1$ ακμές όπου n ο αριθμός των κλικων. Η παρακάτω διαδικασία περιγράφεται στην αναφορά [55]. Καταρχάς ορίζεται το *Συνδετικό γράφημα (Junction graph)* το οποίο έχει σαν κόμβους τις κλίκες και για κάθε ζεύγος κλικών U, V με τομή $S = U \cap V \neq \emptyset$ έχουμε μια σύνδεση με βάρος $|S|$.

Θεώρημα 6.1. *Ένα σκελετικό δέντρο για ένα συνδετικό γράφημα είναι συνδετικό δέντρο αν και μόνο αν είναι σκελετικό δέντρο μέγιστου βάρους.*

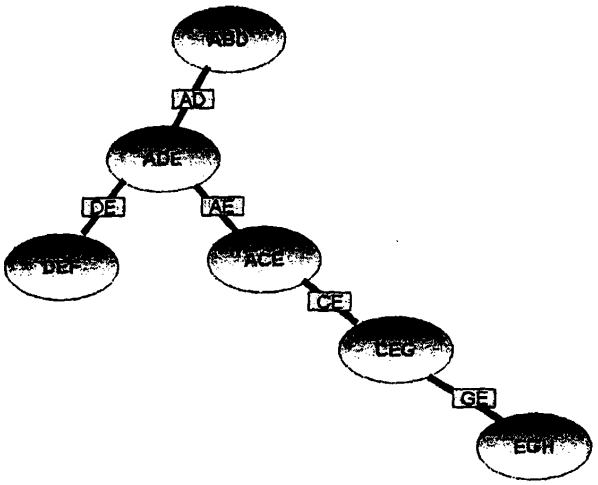
Το θεώρημα 6.1 αποδείχθηκε από τον Jensen [56]. Οπότε αρκεί να βρούμε ένα σκελετικό δέντρο μεγίστου βάρους. Αυτό μπορούμε να επιτευχθεί με κάποιον από τους αλγορίθμους Prim ή Kruskal. Εμείς επιλέξαμε τον Kruskal και το περιγράφουμε στην συνέχεια.

Αλγόριθμος Kruskal

Επιλέγεται επαναληπτικά μία σύνδεση μέγιστου βάρους που δεν δημιουργεί κύκλο. Ο αλγόριθμος του Kruskal δουλεύει με ένα δάσος από μερικώς σκελετικά δέντρα μεγίστου βάρους. Όταν μια σύνδεση επιλέγεται, δύο επιμέρους δέντρα συνδέονται σε ένα νέο μερικώς σκελετικό δέντρο μεγίστου βάρους. Αυτός ο αλγόριθμος όσο και του Prim, έχουν σαν αποτέλεσμα ένα σκελετικό δέντρο μεγίστου βάρους (Goudran and Minoux [57]).

6.2.7 Διαχωριστικά σύνολα (seperator sets) και Δυναμικά (potentials)

Αφού έχει κατασκευαστεί το συνδετικό δέντρο ανάμεσα σε γειτονικές κλίκες βρίσκουμε το διαχωριστικό σύνολο (seperator set) που περιέχει τις κοινές μεταβλητές των δύο γειτονικών



Σχήμα 6.5: Συνδετικό δέντρο

κλικών. Κάθε κλίκα και διαχωριστικό σύνολο θα περιέχει την από κοινού κατανομή των μεταβλητών που περιέχει. Τις κατανομές αυτές ονομάζουμε δυναμικά (potentials).

6.3 - Αρχιτεκτονικές ανταλλαγής μηνυμάτων

Υπάρχουν δύο γνωστές αρχιτεκτονικές για συμπερασματολογία σε Συνδετικά δέντρα(Lauritzen και Spiegelhalter [50], Shenoy και Shafer [58]). Οι διαφορές τους είναι ελάχιστες. Συνήθως χρησιμοποιείται η Lauritzen και Spiegelhalter που είναι και ελαφρώς πιο γρήγορη. Βασίζονται σε ανταλλαγή μηνυμάτων μεταξύ γειτονικών κλικών. Πριν περιγράψουμε την διαδικασία πρέπει να οριστούν οι βασικές πράξεις μεταξύ των δυναμικών (potentials) και τα μηνύματα μεταξύ των κλικών. Τα δυναμικά περιγράφουν την από κοινού κατανομή των μεταβλητών κάθε κλίκας. Στην συνέχεια του κεφαλαίου τα δυναμικά συμβολίζονται με ϕ και δεν πρέπει να τα μπερδέψουμε με τις παραμέτρους τις οποίες συμβολίζαμε με το ίδιο σύμβολο στα προηγούμενα κεφάλαια.

Ορίζονται τρεις βασικές πράξεις

1. Η περιθωριοποίηση ενός δυναμικού ϕ_X ως προς ένα σύνολο Y

$$\sum_{X \setminus Y} \phi_X \tag{6.2}$$

2. Ο πολλαπλασιασμός μεταξύ δύο δυναμικών.

3. Η διαίρεση μεταξύ δύο δυναμικών.

Ανάλογα και με το είδος της κατανομής που έχουμε αυτές οι πράξεις ορίζονται διαφορετικά όπως θα δούμε στην συνέχεια.

Ορισμός 6.2. Ανταλλαγή Μηνύματος

Αν θεωρηθούν δύο γειτονικές κλίκες X και Y με δυναμικά ϕ_X και ϕ_Y και διαχωριστικό σύνολο ϕ_R τότε το μήνυμα από την κλίκα X στην Y αποτελείται από δύο βήματα. Τίθεται ένα νέο δυναμικό στο R, αποθηκεύοντας το παλιό (Projection).

$$\phi_R^{old} \leftarrow \phi_R \tag{6.3}$$

και

$$\phi_R \leftarrow \sum_{X \setminus R} \phi_X \tag{6.4}$$

Τίθεται ένα νέο δυναμικό στο Y (Absorption)

$$\phi_Y \leftarrow \phi_R \frac{\phi_R}{\phi_R^{old}} \tag{6.5}$$

6.3.1 Περιγραφή αλγορίθμου ανταλλαγής μηνυμάτων

Τα βασικά βήματα του αλγορίθμου είναι

- Αρχικοποίηση
- Εισαγωγή Πληροφορίας
- Συλλογή Πληροφορίας (Collect)
- Κατανομή Πληροφορίας (Distribute)

Αρχικοποίση

Κάθε μεταβλητή V ανατίθεται στην κλίκα που περιέχει όλους τους πατέρες της. Μετατρέπουμε το δυναμικό της αντίστοιχης κλίκας έστω X ως εξής

$$\psi_X \leftarrow \psi_X P(V|Pa_V) \tag{6.6}$$

Εισαγωγή Πληροφορίας

Εισάγονται τα στιγμιότυπα για τις μεταβλητές που είναι γνωστά. Εξαρτάται από την κατανομή και θα το περιγράψουμε στις Παραγράφους (6.3.2, 6.3.3).

Συλλογή Πληροφορίας

Κάθε κόμβος (κλίκα) αφού έχει λάβει μηνύματα από όλα τα παιδιά του στέλνει μήνυμα στον πατέρα του μέχρι να φτάσουμε στην ρίζα του δέντρου. Τώρα η ρίζα περιέχει την σωστή από κοινού κατανομή των μεταβλητών της.

Κατανομή Πληροφορίας

Κάθε κόμβος αφού λάβει μήνυμα από τον πατέρα του στέλνει μήνυμα στα παιδιά του μέχρι να ενημερωθούν όλα τα φύλλα. Μετά και το δεύτερο πέρασμα έχουμε ένα *συνεπές* συνδετικό δέντρο . Κάθε δυναμικό κλίκας ή διαχωριστικού συνόλου περιέχει την σωστή από κοινού κατανομή των μεταβλητών της.

Ορθότητα του σχήματος

Γενικά η απόδειξη της ορθότητας είναι αρκετά δύσκολη. Πρώτες αποδείξεις ήταν σε υποπεριπτώσεις Bayesian δικτύων όπως polytrees [34]. Ωστόσο το γεγονός ότι ανατίθεται κάθε παράγοντας P_i σε μία μοναδική κλίκα μας εξασφαλίζει ότι δεν υπολογίζεται παραπάνω από όσο πρέπει πληροφορία. Επίσης, η ιδιότητα των συνδετικών δέντρων, εξασφαλίζει ότι όταν δύο υποδέντρα στέλνουν μήνυμα σε έναν κόμβο, η πληροφορία συνδυάζεται σωστά. Μια απόδειξη του Lauritzen-Spiegelhalter σχήματος δίνεται στην εργασία [59].

6.3.2 Βασικές πράξεις στις πολυνυμικές (multinomial) κατανομές

Τα δυναμικά είναι πίνακες που περιέχουν την πιθανότητα για το αντίστοιχη ανάθεση τιμών του συνόλου μεταβλητών τους.

Εισαγωγή πληροφορίας

Έστω ότι έχουμε την παρατήρηση ότι μια μεταβλητή V παίρνει την τιμή v. Για να εισαχθεί αυτή η πληροφορία στο συνδετικό δέντρο αρκεί σε μία κλίκα που περιέχει την μεταβλητή V να μηδενιστούν οι τιμές στο δυναμικό της κλίκας όπου $V \neq v$. Κατά την φάση της συλλογής και της κατανομής αυτή η πληροφορία διαδίδεται στο δέντρο, για αυτό δεν χρειάζεται να εισαχθεί η πληροφορία σε όλες τις κλίκες από την αρχή.

Περιθωριοποίηση

Έστω ο πίνακας που περιέχει την από κοινού κατανομή των διακριτών μεταβλητών (X, Y, Z). Τότε η περιθωριοποίηση ως προς την μεταβλητή Z ορίζεται ως εξής

$$P(X,Y) = \sum_{(X,Y,Z)\setminus Z} P(X,Y,Z) = \sum_{i=1}^{r_Z} P(X,Y,Z=z_i)$$

όπου r_Z ο αριθμός των πιθανών τιμών του Z και z_i η i-οστή πιθανή τιμή.

Πολλαπλασιαμός και Διαίρεση

Έστω δύο σύνολα διακριτών μεταβλητών X και Y με αντίστοιχα δυναμικά ψ_X και ψ_Y. Ορίζεται σαν πολλαπλασιασμό το δυναμικό με σύνολο μεταβλητών $V = X \cup Y$ όπου $V_i = X_{x=i} * Y_{y=i}$ και $i = 1...r_v$. Η διαίρεση γίνεται αντίστοιχα με τον πολλαπλασιασμό.

Συρρίκνωση της πληροφορίας (Evidence Shrinking)

Η συρρίκνωση της πληροφορίας είναι μια βελτιστοποίηση των Συνδετικών δέντρων με διακριτές μεταβλητές ώστε να μειωθεί η διάσταση των κλικών κατά την εισαγωγή της πληροφορίας. Για παράδειγμα ας υποθέσουμε ότι έχουμε μια κλίκα C με δυαδικές διακριτές μεταβλητές X, Y και Z. Η διάσταση του δυναμικού είναι $2^3 = 8$. Υποθέτουμε τώρα ότι έχουμε την πληροφορία $X = x_1$ και $Y = y_0$. Αν μηδενιστούν οι αντίστοιχες τιμές στο δυναμικό τότε μόνο δύο τιμές θα είναι διάφορες του μηδέν. Οι υπόλοιπες απλά θα διαδοθούν στο δέντρο χωρίς να επηρεάσουν το αποτέλεσμα, απλά επιβαρύνωντας το υπολογιστικό κόστος του αλγορίθμου. Αυτό που γίνεται είναι η συρρίκνωση του δυναμικού ώστε το μέγεθος του να μειωθεί στο πραγματικά ωφέλιμο, στο παράδειγμα μας αντί για 8, γίνεται 2. Η διαφορά είναι ότι πρέπει να εισαχθεί η πληροφορία σε κάθε κλίκα που περιέχει τις μεταβλητές για τις οποίες έχουμε πληροφορία, ώστε να προκύψει ένα συνδετικό δέντρο με όσο πιο δυνατό μειωμένες κλίκες. Στην ουσία κάθε μεταβλητή για την οποία έχουμε πληροφορία αφαιρείται από το δέντρο, μειώνοντας έτσι την διάσταση του προβλήματος [60].

6.3.3 Βασικές πράξεις στην Gaussian περίπτωση

Κανονικά Χαρακτηρηστικά

Έστω ότι έχουμε για κάθε κόμβο V στην γενική περίπτωση μια υπό συνθήκη κατανομή, σύμφωνα με την (4.13) έχουμε:

$$f(x|z) = c \cdot \exp[-\frac{1}{2}((x - \vec{\mu} - B^T z)^T \Sigma^{-1}(x - \vec{\mu} - B^T z)], \tag{6.7}$$

Για να οριστούν οι βασικές πράξεις πρέπει να μετατραπεί σε κανονική μορφή. Η κανονική μορφή για μια Gaussian κατανομή γράφεται γενικά ως:

$$f(x) = c \cdot \exp[-\frac{1}{2}(x^T K x + h^T x + g], \tag{6.8}$$

όπου K, h και g ονομάζονται κανονικά χαρακτηριστικά. Η σχέση (6.7) μπορεί να γραφεί ως:

$$f(x|z) = exp[-\frac{1}{2}\begin{pmatrix} x & z \end{pmatrix}\begin{pmatrix} \Sigma^{-1} & -\Sigma^{-1}B^T \\ -B\Sigma^{-1} & B\Sigma^{-1}B^T \end{pmatrix}\begin{pmatrix} x \\ z \end{pmatrix}$$
$$+\begin{pmatrix} x & z \end{pmatrix}\begin{pmatrix} \Sigma^{-1}\vec{\mu} \\ -B\Sigma^{-1}\vec{\mu} \end{pmatrix} - \frac{1}{2}\vec{\mu}^T\Sigma^{-1}\vec{\mu} + \log c], \tag{6.9}$$

όπου $c = (2\pi)^{-n/2}|\Sigma^{-1}|^{-\frac{1}{2}}$. Οπότε προκύπτει το σύνολο των κανονικών χαρακτηριστικών

$$g = -\frac{1}{2}\vec{\mu}^T\Sigma^{-1}\vec{\mu} + \log c, \tag{6.10}$$

$$h = \begin{pmatrix} \Sigma^{-1}\vec{\mu} \\ -B\Sigma^{-1}\vec{\mu} \end{pmatrix}, \tag{6.11}$$

$$K = \begin{pmatrix} \Sigma^{-1} & -\Sigma^{-1}B^T \\ -B\Sigma^{-1} & B\Sigma^{-1}B^T \end{pmatrix}. \tag{6.12}$$

Εισαγωγή των δεδομένων

Έστω η παρατήρηση ότι μια μεταβλητή Y παίρνει την τιμή y. Αυτό έχει σαν αποτέλεσμα να μειωθεί η διάσταση του χώρου των μεταβλητών και πρέπει να ενημερώσουμε κάθε δυναμικό που περιέχει την Y. Το νέο δυναμικό είναι

$$\phi^*(x) = \exp[g + h_y^T y - \frac{1}{2} y^T K_{YY} y) + x^T (h_x - K_{XX} y) - \frac{1}{2} x^T K_{XX} x]. \tag{6.13}$$

Επομένως μπορούμε να ορίσουμε τα νέα κανονικά χαρακτηριστικά, $\hat{g} = (g + h_y^T y - \frac{1}{2} y^T K_{YY} y)$, $\hat{h} = (h_x - K_{XX} y)$ και $\hat{K} = K_{XX}$.

Πολλαπλασιασμός και διαίρεση

Πρώτα επεκτείνονται τα δύο δυναμικά στον ίδιο χώρο μεταβλητών προσθέτοντας στις κατάλληλες διαστάσεις μηδενικά και υπολογίζεται για τον πολλαπλασιασμό

$$(g_1, h_1, K_1) * (g_2, h_2, K_2) = (g_1 + g_2, h_1 + h_2, K_1 + K_2),$$

και αντίστοιχα με $-$ για την διαίρεση.

Περιθωριοποίηση

Έστω ένα δυναμικό ϕ_W σε ένα σύνολο μεταβλητών W. Μπορούμε να υπολογίσουμε το δυναμικό ενός υποσυνόλου $V \subset W$ με περιθωριοποίηση, που δηλώνεται ως $\phi_V = \sum_{W \setminus V} \phi_W$. Έστω $y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$, $h = \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}$ και $K = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix}$, μπορεί να δειχθεί ότι

$$\int \phi[(y_1^T y_2^T)^T] dy_1 = \phi[y_2; \hat{g}, \hat{h}, \hat{K}], \tag{6.14}$$

όπου

$$\hat{g} = g + \frac{1}{2}(p \log(2\pi) - \log|K_{11}| + h_1^T K_{11}^T h_1), \tag{6.15}$$

$$\hat{h} = h_2 - K_{21} K_{11}^{-1} h_1, \tag{6.16}$$

$$\hat{K} = K_{22} - K_{21} K_{11}^{-1} K_{12}. \tag{6.17}$$

Εύρεση δυναμικό μεταβλητών που είναι διάσπαρτες στο δέντρο

Στον αλγόριθμο αναζήτησης όπως περιγράψαμε σε κάθε επανάληψη δημιουργείται ένα νέο δίκτυο με μία μόνο αλλαγή στο τρέχων δίκτυο. Αυτή η αλλαγή μπορεί να είναι είτε πρόσθεση μίας ακμής, είτε αφαίρεση είτε εναλλαγή. Αυτές οι πράξεις έχουν σαν αποτέλεσμα την αλλαγή ενός ή το πολύ δύο τοπικών κατανομών. Για να αξιολογηθεί το νέο δίκτυο που προκύπτει πρέπει να βρεθούν τα αναμενόμενα επαρκή στατιστικά της νέας ή των νέων κατανομών που προκύπτουν. Αυτό επιτυγχάνεται με κάποιες αλλαγές στο συνδετικό δέντρο. Η γενική ιδέα είναι να αντικατασταθούν οι ελλιπείς τιμές με τις αναμενόμενες τιμές κάθε

κατανομής που προκύπτουν από το συνεπές συνδετικό δέντρο. Αν η νέα κατανομή είναι υποσύνολο της παλιάς (αυτό συμβαίνει με την διαγραφή κάποιου πατέρα), τότε τα πράγματα είναι σχετικά απλά. Στο δυναμικό του δέντρου που περιέχει την κατανομή του κόμβου, γίνεται περιθωριοποίηση στις μεταβλητές που δεν εμφανίζονται στην νέα κατανομή. Όταν οι μεταβλητές είναι διάσπαρτες σε διάφορες κλίκες τα βήματα που ακολουθούνται είναι τα εξής:

- Χρησιμοποιείται το ελάχιστο υποδέντρο που περιέχει όλες τις μεταβλητές ενδιαφέροντος, δηλαδή αυτές που περιέχονται στην νέα κατανομή.
- Για κάθε φύλλο του νέου δέντρου, γίνεται περιθωριοποίηση στις μεταβλητές που δεν περιέχονται σε αυτές του ενδιαφέροντος ούτε στο διαχωριστικό σύνολο του κόμβου. Στην συνέχεια αν ϕ το δυναμικό του φύλλου και ψ της κλίκας πιο κοντά στην ρίζα, το νέο δυναμικό ψ' της κλίκας που είναι πιο κοντά στη ρίζα προκύπτει ως:

$$\psi' = \frac{\psi \cdot \phi}{\psi_S}, \tag{6.18}$$

  όπου ψ_S είναι το δυναμικό του διαχωριστικού συνόλου.
- Αφαιρούνται τα φύλλα από το δέντρο και επαναλαμβάνεται η διαδικασία έως ότου η ρίζα να περιέχει μόνο τις μεταβλητές ενδιαφέροντος.

Το δυναμικό που προκύπτει από την παραπάνω διαδικασία έχει τις αναμενόμενες τιμές της νέας κατανομής ως προς το τρέχον δίκτυο.

6.4 Υβριδική περίπτωση

Παρακάτω θα παρουσιάσουμε το πρόβλημα της συμπερασματολογίας σε υβριδικά δίκτυα δηλαδή σε δίκτυα που περιέχουν τόσο διακριτές όσο και συνεχείς μεταβλητές. Πριν περάσουμε στην περιγραφή πρέπει να οριστούν δύο νέες έννοιες

Ορισμός 6.3. Ισχυρά Συνδετικά δέντρα (Strong Junction Trees) Αν Συνδετικό δέντρο με κλίκες C και ρίζα $\mathcal{R} \in C$ τέτοιο ώστε για όλες τις γειτονικές κλίκες, C και D με C κοντινότερα στην ρίζα, τότε ισχύει

$$S = C \cap D \subseteq \Delta, \tag{6.19}$$

ή

$$D \backslash C \subseteq \Gamma, \tag{6.20}$$

όπου Δ το σύνολο των διακριτών μεταβλητών και Γ των συνεχών.

Ορισμός 6.4. Υβριδικά δυναμικά Οι κατανομές αυτές περιέχουν τόσο συνεχείς όσο και διακριτές μεταβλητές. Για κάθε πιθανή ανάθεση τιμής i των διακριτών μεταβλητών Y, η κατανομή των συνεχών μεταβλητών X είναι μια Gaussian κατανομή που έχει την μορφή

$$f(X|Y=i) = N(X, \mu(i), \Sigma(i)), \tag{6.21}$$

όπου $\mu(i)$ και $\Sigma(i)$ το μέσο και ο πίνακας συμμεταβλητότητας στην i-οστή πιθανή ανάθεση τιμών των διακριτών μεταβλητών Y.

Υπάρχουν δύο γνωστοί αλγόριθμοι που βασίζονται σε Συνδετικά δέντρα για την υβριδική περίπτωση. Ο πρώτος αλγόριθμος [61], είναι στην ουσία είναι επέκταση της συνεχούς περίπτωσης που είδαμε παραπάνω. Σε αυτή έχουμε ένα σύνολο κανονικών χαρακτηριστικών για κάθε διακριτή τιμή. Η μόνη πράξη που έχει διαφορά από τα παραπάνω είναι η περιθωριοποίηση για μια διακριτή τιμή όπου τα μέσα και οι πίνακες συμμεταβλητότητας του δυναμικού εξαρτώνται από αυτήν την μεταβλητή, δηλαδή κάποια συνεχείς μεταβλητή της κλίκας έχει πατέρα στο γράφημα την συγκεκριμένη διακριτή μεταβλητή. Σε αυτήν την περίπτωση προκύπτει μια μεικτή Gaussian κατανομή:

$$\sum_j \phi(x,j,i) = \sum_j p \times Q(x;\vec{\mu(i)},\Sigma(i)). \tag{6.22}$$

Οπότε πρέπει να κρατηθεί μία λίστα από όρους. Η περιθωριοποίηση όλων των συνεχών μεταβλητών θα πρέπει να γίνεται πριν από αυτή των διακριτών από τις οποίες εξαρτώνται. Αυτό μπορεί να επιτευχθεί αν κατά την κατασκεύη του Συνδετικού δέντρου διαγραφούν πρώτα οι συνεχείς μεταβλητές και στην συνέχεια οι διακριτές. Μια τέτοια σειρά διαγραφής ονομάζεται ισχυρή τριγωνοποίηση και οδηγεί σε ισχυρά συνδετικά δέντρα. Ωστόσο δεν μπορεί να αποφευχθεί η παραπάνω περίπτωση για την εύρεση της περιθωριοποίησης μεταβλητών που είναι υποσύνολο μιας κλίκας ή μεταβλητών που δεν ανήκουν σε μία μόνο κλίκα. Η κλασική προσέγγιση είναι η συγχωνεύση της μεικτής Gaussian κατανομής σε k στοιχεία. Για $k=1$ προκύπτει:

$$\hat{p}(i) = \sum_j p(i,j), \tag{6.23}$$

$$\hat{\vec{\mu}}(i) = \sum_j \vec{\mu}(i,j)p(i,j)/\hat{p(i)}, \tag{6.24}$$

$$\hat{\Sigma}(i) = \sum_j \Sigma(i,j)p(i,j)/\hat{p(i)} + \sum_j (\vec{\mu}(i,j)-\hat{\vec{\mu}}(i))(\vec{\mu}(i,j)-\hat{\vec{\mu}}(i))^T p(i,j)/\hat{p(i)} \tag{6.25}$$

Η οποία είναι η καλύτερη προσέγγιση με βάση την απόσταση Kullback (KL) [62]. Στην παραπάνω προσέγγιση έχουν εντοπιστεί κάποιες αριθμητικές αστάθειες που οφείλονται κυρίως στην μετατροπή σε κανονικά χαρακτηρηστικά και αντίστροφα των Gaussian κατανομών. Για την επίλυση αυτού του προβλήματος οι Lauritzen και Jensen [63] πρότειναν ένα νέο σχήμα που περιγράφεται παρακάτω.

6.5 Τοπικοί σταθεροί υπολογισμοί (Stable Local Computations)

Για την επίλυση του προβλήματος των αριθμητικών ασταθειών της προηγούμενης προσέγγισης οι Lauritzen και Jensen [63] ανέπτυξαν ένα νέο σχήμα που βασίζεται στο κλασικό, της διάδοσης μηνυμάτων (Lauritzen και Spiegelhalter) αλλά τα δυναμικό και τα μηνύματα που χρησιμοποιούν είναι αρκετά διαφορετικά. Πρώτα περιγράφουμε τα νέα δυναμικό και στην συνέχεια τις βασικές λειτουργίες του σχήματος.

6.5.1 Υπό συνθήκες Gaussian (CG) δυναμικά

Στον αλγόριθμο που περιγράφεται στην συνέχεια χρησιμοποιούνται οι υπό συνθήκες Gaussian κατανομές (CG) που πρωτοπαρουσιάστηκαν από τους Lauritzen και Wermuth [64]. Θα τις περιγράψουμε εν συντομία. Περισσότερες πληροφορίες παραθέτονται στην αναφορά [61]. Ένα CG δυναμικό αναπαριστάται ως $\phi = [p, A, B, C](H|T)$, όπου με $(H|T)$ δηλώνεται η διαμέριση του συνόλου των συνεχών μεταβλητών σε κεφαλή και ουρά. Δηλώνονται οι μεταβλητές της κεφαλής με Y και της ουράς με Z με διαστάσεις r και s αντίστοιχα. Ένα οποιοδήποτε σύνολο πιθανών τιμών για τις διακριτές μεταβλητές του δυναμικό δηλώνεται με i. Έτσι κάθε δυναμικό, έχει ένα σύνολο συνεχών μεταβλητών, κεφαλής και ουρά, που κάποιο από αυτά μπορεί να είναι κενό. Στην παραπάνω αναπαράσταση έχουμε

- $p = \{p(i)\}$ η πιθανότητα της διακριτής τιμής i όπως και στο «συνηθισμένο» διακριτό δυναμικό.
- $A = \{A(i)\}$ ένας πίνακας από $r \times 1$ διανύσματα.
- $B = \{B(i)\}$ ένα σύνολο από $r \times s$ πίνακες.
- $C = \{C(i)\}$ ένα σύνολο από $r \times r$ θετικά ημιορισμένους συμμετρικούς πίνακες.

Το δυναμικό που αναπαριστάται από $\phi = [p, A, B, C](H|T)$ καθορίζει την CG παλινδρόμηση

$$P(I = i) \propto p(i), \quad \mathcal{L}(Y|I = i, Z = z) = N_r(A(i) + B(i)z, C(i)) \tag{6.26}$$

Επέκταση και μείωση

Ένα CG δυναμικό μπορεί να επεκταθεί προσθέτοντας διακριτές ή συνεχείς μεταβλητές στην ουρά του. Όταν προστίθενται διακριτές μεταβλητές, απλά τίθεται $p^*(i, j) = p(i)$. Όταν προστίθενται συνεχείς μεταβλητές στην ουρά, επεκτείνεται ο B πίνακας προσθέτοντας μηδενικά στις στήλες που αντιστοιχούν στις νέες μεταβλητές της ουράς, οπότε προκύπτει ένας νέος πίνακας

$$B^* = \{B(i) : 0\}. \tag{6.27}$$

Όμοια, αν ο B έχει στήλες που έχουν μόνο μηδενικά τότε μπορούν να αφαιρεθούν οι μεταβλητές της ουράς που αντιστοιχούν στις στήλες με τα μηδενικά. Το δυναμικό που προκύπτει ονομάζεται *ελάχιστο*.

Περιθωριοποίηση

Όπως και στο αρχικό σχήμα η περιθωριοποίηση ενός δυναμικού ορίζεται σε μερικές μόνο περιπτώσεις και επίσης οι περιθωριοποίησεις πάνω στις συνεχείς μεταβλητές υπολογίζονται πριν τις περιθωριοποίησεις πάνω στις διακριτές. Οι περιθωριοποιήσεις πάνω στις συνεχείς ορίζονται μόνο σε μεταβλητές της κεφαλής του δυναμικού. Το δυναμικό $[p, A, B, C](H, T)$ «σπάει» ως εξής

$$H = (H_1, H_2), A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}, B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}, C = \begin{pmatrix} C_{11} & C12 \\ C_{21} & C_{22} \end{pmatrix} \tag{6.28}$$

που αντιστοιχούν στην διαμέριση της κεφαλής σε δύο υποσύνολα μεταβλητών. Η περιθωριοποίηση του ϕ ορίζεται ως εξής:

$$\phi^{\downarrow D'} = [p, A_1, B_1, C_{11}](H_1|T) \tag{6.29}$$

Αυτές οι περιθωριοποιήσεις ονομάζονται *ισχυρές* καθώς αντιστοιχούν στον υπολογισμό των κανονικών περιθωριοποιήσεων των αντίστοιχων υπό συνθήκη Gaussian κατανομών. Όταν όλες οι μεταβλητές από την κεφαλή έχουν αφαιρεθεί τότε μπορούν να αφαιρεθούν οι μεταβλητές ουράς και προκύπτει ένα συνηθισμένο διακριτό δυναμικό.

Οι περιθωριοποιήσεις πάνω σε διακριτές μεταβλητές ορίζονται μόνο αν η ουρά είναι άδεια, οπότε δεν έχουμε και τον πίνακα B. Τότε η περιθωριοποίηση του CG δυναμικού $\phi = [p, A, -, C](H|-)$, με σύνολο διακριτών μεταβλητών διαμερισμένο σε $U \cup W$, πάνω στο W ορίζεται ως εξής:

$$\phi^{\downarrow U \cup H} = [\tilde{p}, \tilde{A}, -, \tilde{C}](H|-) \tag{6.30}$$

όπου

$$\tilde{p}(i_U) = \sum_W, \tag{6.31}$$

$$\tilde{A}(i_U) = \frac{1}{\tilde{p}(i_U)} \sum_W A(i)p(i), \tag{6.32}$$

$$\tilde{C}(i_U) = \frac{1}{\tilde{p}(i_U)} \sum_W C(i) + [A(i) - \tilde{A}(i_U)][A(i) - \tilde{A}(i_U)]^T p(i), \tag{6.33}$$

όπου $i = (i_U, i_W)$. Αυτή η περιθωριοποίηση, όπως ήδη έχουμε αναφέρει, ονομάζεται ασθενής και είναι μια προσέγγιση της μεικτής Gaussian κατανομής.

Άμεσος συνδυασμός

Ο άμεσος συνδυασμός ορίζεται για συγκεκριμένα ζευγάρια δυναμικών και αυτό διαφοροποιεί το σχήμα από τα υπάρχοντα. Ο άμεσος συνδυασμός δύο δυναμικών $\phi = [p, A, B, C](H_1|T_1)$ και $\psi = [q, E, F, G](H_2|T_2)$ ορίζεται μόνο αν η τομή της κεφαλής του δεύτερου δυναμικό με το σύνολο των συνεχών μεταβλητών του πρώτου είναι κενό σύνολο:

$$H_2 \cap D_1 = \emptyset \tag{6.34}$$

Εδώ πάντα υποθέτουμε ότι τα δυναμικό είναι ελάχιστα. Αν η συνθήκη (6.34) ικανοποιείται, τότε επεκτείνονται τα δυναμικά ώστε για τις επεκτάσεις να ισχύει $T_2 = H_1 \cup T_1$. Αυτό επιτυγχάνεται επεκτείνοντας την ουρά T_1 με $T_1 \cup (T_2 \backslash H_1)$ και την ουρά T_2 με $T_2 \cup H_1 \cup T_1$. Στην συνέχεια, διαμερίζεται το F σε $F = [F_1 : F_2]$ όπου F_1 $(r_2 \times r_1)$ που αντιστοιχεί στο H_1 και F_2 $(r_2 \times s_1)$ που αντιστοιχεί στο T_1. Τότε ορίζεται ο άμεσος συνδυασμός ως

$$[\rho, U, V, W](H|T) = [p, A, B, C](H_1|T_1) \dot{\otimes} [q, E, F, G](H_2|T_2) \tag{6.35}$$

$$\rho = pq, \tag{6.36}$$

$$U = \begin{pmatrix} A \\ E + F_1 A \end{pmatrix}, \tag{6.37}$$

$$V = \begin{pmatrix} B \\ F_2 + F_1 B \end{pmatrix}, \tag{6.38}$$

$$W = \begin{pmatrix} C & CF_1^T \\ F_1 C & G + F_1 C F_1^T \end{pmatrix}. \tag{6.39}$$

Αυτός ο συνδυασμός αντιστοιχεί στην συνηθισμένη σύνθεση υπό συνθήκη κατανομών. Σημειώνουμε ότι αν υπάρχουν τα $\phi \dot{\otimes} \psi$ και $\psi \dot{\otimes} \phi$ τότε είναι ισοδύναμα. Επίσης ο άμεσος συνδυασμός ικανοποιεί την ιδιότητα

$$(\phi \dot{\otimes} \psi) \dot{\otimes} \eta = \phi \dot{\otimes} (\phi \dot{\otimes} \eta) \tag{6.40}$$

με την έννοια ότι αν ορίζονται και οι δύο συνδυασμοί, το αποτέλεσμα θα είναι το ίδιο [65].

Συμπληρώματα

Αν η κεφαλή ενός CG δυναμικό $\phi = [p, A, B, C](H|T)$ διαμερίζεται ως

$$H = (H_1, H_2), A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}, B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}, C = \begin{pmatrix} C_{11} & C12 \\ C_{21} & C_{22} \end{pmatrix} \tag{6.41}$$

και $[p^*, A_1, B_1, C_{11}](H_1|T)$ είναι η ισχυρή περιθωριοποίηση του ϕ, τότε ορίζεται το συμπλήρωμα $\phi^{|H_1 \cup T}$ ως $[q, E, F, G](H_2|H_1 \cup T)$ όπου

$$q = p/p^*, \tag{6.42}$$

$$E = A_2 - C_{21} C_{11}^- A_1, \tag{6.43}$$

$$F = [C_{21} C_{11}^- : B_2 - C_{21} C_{11}^- B_1], \tag{6.44}$$

$$G = C_{22} - C_{21} C_{11}^- C_{12}. \tag{6.45}$$

όπου C_{11}^- ο ψευδό-αντίστροφος του C_{11}.

Ισχύει

$$[p, A, B, C](H, T) = [p^*, A_1, B_1, C_{11}](H_1|T) \dot{\otimes} [q, E, F, F](H_2|H_1 \cup T). \tag{6.46}$$

Η αποσύνθεση του δυναμικό σε περιθωριοποιημένο και συμπλήρωμα αντιστοιχεί επακριβώς στην αποσύνθεση της κατανομής πιθανότητας στην περιθωριοποιημένη και στην υπό συνθήκη κατανομή.

Αναδρομικός συνδυασμός

Στην συνέχεια ορίζεται ένας πιο γενικός συνδυασμός των CG δυναμικών. Αυτό απαιτείται για την αρχικοποίηση που περιγράφεται στην συνέχεια. Θεωρούμε πάλι δύο δυναμικά $\phi =$

$[p, A, B, C](H_1|T_1)$ και $\psi = [q, E, F, G](H_2|T_2)$. Αν $H_1 \cap H_2 \neq \emptyset$ ο συνδυασμός δεν ορίζεται προς το παρόν. Αν η τομή είναι κενή, ο συνδυασμός ορίζεται ως

$$\phi \otimes \psi = \psi \dot{\otimes} \phi \text{ ή } \phi \dot{\otimes} \psi = \psi \otimes \phi. \tag{6.47}$$

αν ορίζεται έστω ένας από τους δύο άμεσους συνδυασμούς των δεξιών μελών. Αν ορίζονται και οι δύο, οι συνδυασμοί είναι ισοδύναμοι. Αν κανένα από τα δύο δεν ορίζεται, πρέπει να ισχύει

$$H_1 \cap D_2 \neq \emptyset \text{ και } H_2 \cap D_1 \neq \emptyset. \tag{6.48}$$

Έστω $D_{12} = H_1 \backslash D_2$ και $D_{21} = H_2 \backslash D_1$. Αν και τα δύο είναι κενά τότε ο συνδυασμός δεν ορίζεται. Αλλιώς αποσυντίθεται κάποιος παράγοντας, π.χ. ο ϕ αν υποτεθεί ότι $D_{12} \neq \emptyset$,

$$\phi = \phi^{\downarrow(D_1 \backslash D_{12})} \dot{\otimes} \phi^{|(D_1 \backslash D_{12})} = \phi' \dot{\otimes} \phi'', \tag{6.49}$$

και γίνεται προσπάθεια να συνδυαστούν τα ϕ και ψ ως εξής

$$\phi \otimes \psi = (\phi' \otimes \psi) \dot{\otimes} \phi''. \tag{6.50}$$

Αυτή η εξίσωση μπορεί να θεωρηθεί αναδρομική με την έννοια ότι η διαδικασία που περιγράφηκε επαναλαμβάνεται για το $(\phi' \otimes \psi)$, ενώ ο άμεσος συνδυασμός ορίζεται από την κατασκευή.

Συλλογής πληροφορίας

Όταν ένα μήνυμα στην φάση συλλογής στέλνεται από μία κλίκα C σε μία γειτονική της D προς την ρίζα, με διαχωριστικό σύνολο $S = C \cap D$, τα δυναμικά ϕ_C στον C και ϕ_D στην D αλλάζουν και γίνονται ϕ_C^* και ϕ_D^*, με

$$\phi_C^* = \phi_C^{|S},\ \phi_D^* = \phi_D \otimes \phi_C^{\downarrow S} \tag{6.51}$$

όπου ϕ_C^* το συμπλήρωμα του ϕ_C μετά από περιθωριοποίηση στο διαχωριστικό σύνολο και ϕ_D^* με συνδυασμό του αρχικού δυναμικό με την περιθωριοποίηση του ϕ_C. Η απόδειξη ότι ο συνδυασμός είναι καλά ορισμένος δίνεται στην αναφορά [63].

Κατανομή πληροφορίας

Όταν ένα μήνυμα στην φάση κατανομής στέλνεται από μία κλίκα C σε μία γειτονική της D μακρύτερα από την ρίζα με διαχωριστικό σύνολο $S = C \cap D$, έχει ήδη λάβει μήνυμα από τον γείτονα του προς την ρίζα. Κάνουμε την υπόθεση ότι το διαχωριστικό δυναμικό S' προς την ρίζα περιέχει την ασθενή περιθωριοποίηση του δυναμικού

$$\phi_{S'} = \phi^{\downarrow S'}. \tag{6.52}$$

Όταν στέλνεται ένα μήνυμα, ένα νέο δυναμικό δημιουργείται ϕ_S στο S. Πρώτα υπολογίζεται η ασθενής περιθωριοποίηση της κλίκας C

$$\phi^{\downarrow C} = \phi_{S'} \dot{\otimes} \phi_C. \tag{6.53}$$

Στην συνέχεια υπολογίζουμε την περιθωριοποίηση ως προς το διαχωριστικό σύνολο

$$\phi_S = (\phi^{1C})^{1S} = \phi^{1S}. \tag{6.54}$$

Ο συνδυασμός είναι καλά ορισμένος επειδή μετά την φάση συλλογής, τα συμπληρωματικά δυναμικά αποθηκεύτηκαν στις κλίκες, άρα η τομή της κεφαλής του ϕ_C και του S' είναι κενό σύνολο, και οι (ασθενείς) περιθωριοποιήσεις είναι καλά ορισμένες αφού η ουρά του ϕ_C περιέχεται στην κεφαλή του $\phi_{S'}$ που έχει σαν συνέπεια ο συνδυασμός (6.53) να έχει κενή ουρά. Παρόλο που στο [63] επιλέγεται να μην αποθηκεύουμε τις (πιθανώς ασθενείς) περιθωριοποιήσεις στα δυναμικά των κλικών, εμείς επιλέγουμε να τα αποθηκεύουμε γιατί γίνεται πιο εύκολη η διαχείριση του πληροφορίας όπως θα δούμε στα συνέχεια.

Εισαγωγή Πληροφορίας

Η διαχείριση της πληροφορίας γίνεται μετά την αρχικοποίηση (φάση συλλογής). Η διαχείριση της διακριτής πληροφορίας γίνεται όπως συνήθως και χρειάζεται να εισαχθεί η πληροφορία για κάθε μεταβλητή μόνο σε μία κλίκα που περιέχει την αντίστοιχη μεταβλητή . Για τις συνεχείς μεταβλητές τα πράγματα είναι κάπως διαφορετικά. Πρώτα η πληροφορία πρέπει να εισαχθεί σε όλες τις κλίκες που περιέχεται η μεταβλητή. Επίσης στο [63] έχουμε τον διαχωρισμό, αν η μεταβλητή περιέχεται στην κεφαλή ή στην ουρά κάθε δυναμικό. Αλλά όπως επιλέξαμε εμείς να κάνουμε την αρχικοποίηση όλες οι μεταβλητές περιέχονται στις κεφαλές των δυναμικό. Οπότε θα εξετάσουμε μόνο αυτήν την περίπτωση. Η ορθότητα αποδεικνύεται στο [63].

Έστω η μεταβλητή Y_2 εμφανίζεται με τιμή y_2. Εισάγεται σε όλα τα δυναμικό τα οποία μετατρέπονται σε $\phi^* = [p^*, A^*, -, C^*](H^*|-)$, όπου η κεφαλή H^* προκύπτει με αφαίρεση της Y_2 . Διακρίνονται τώρα δύο περιπτώσεις

1. Αν υπάρχει j ώστε $C_{22}(j) = 0$ και $y_2 = A_2(j)$, τίθεται για κάθε i

$$p^*(i) = \left\{ \begin{array}{ll} p(i) & \text{Αν } y_2 = A_2(i) \text{ και } C_{22}(i) = 0 \\ 0 & \text{Αλλιώς} \end{array} \right\}, \tag{6.55}$$

και για κάθε $p^*(i) > 0$ τίθεται

$$A^*(i) = A_1(i),\ C^*(i) = C_{11}(i). \tag{6.56}$$

2. Αλλιώς

$$p^* = \left\{ \begin{array}{ll} \frac{p(i)}{\sqrt{2\pi C_{22}(i)}} \exp[-\frac{1}{2}(y_2 - A_2(i))^2/C_{22}(i)] & \text{Αν } C_{22}(i) \neq 0 \\ 0 & \text{Αλλιώς} \end{array} \right\}, \tag{6.57}$$

και για κάθε $p^*(i) > 0$ τίθεται

$$A^*(i) = A_1(i) + C_{12}(i)(y_2 - A_2(i))/C_{22}(i), \tag{6.58}$$

$$C^*(i) = C_{11}(i) - C_{12}(i)C_{21}(i)/C_{22}(i). \tag{6.59}$$

Κεφαλαιο 7

Περιγραφη Προσεγγισης

7.1 Γενικά

Σε αυτό το κεφάλαιο περιγράφουμε τον παράλληλο αλγόριθμο για την εύρεση του καλύτερου Bayesian δικτύου από τα δεδομένα που αναπτύξαμε. Η σειριακή υλοποίηση δεν είναι αποδοτική για την εύρεση μεγάλων δικτύων που είναι και ο στόχος μας. Σημειώνουμε ότι στη τρέχουσα προσέγγιση δεν εισάγεται εκ των προτέρων γνώση στην δομή του δικτύου. Ο ευρετικός αλγόριθμος είναι κοινός για όλες τις περιπτώσεις και περιγράφεται στην παράγραφο 7.3.

Παράλληλος K-NN για ανάκτηση χαμένων τιμών

Πρώτα προσεγγίζουμε τις ελλιπείς τιμές κάνοντας χρήση του K-NN σε μια παράλληλη μορφή. Αν έχουμε P επεξεργαστές και K ελλιπείς τιμές τότε αναθέτουμε σε κάθε επεξεργαστή K/P ελλιπείς τιμές προς ανάκτηση με τον κλασσικό αλγόριθμο που περιγράψαμε στην παράγραφο 3.2. Η διαδικασία είναι πολύ απλή, σχετικά γρήγορη και έχει μελετηθεί εκτενώς στην βιβλιογραφία [27], οπότε δεν παραθέτουμε πειράματα για αυτήν την διαδικασία. Η ανάκτηση χαμένων τιμών με SVD έχει υλοποιηθεί σειριακά.

Ομαδοποίηση

Αν ο αριθμός των μεταβλητών είναι πολύ μεγάλος μπορεί να μειωθεί, χρησιμοποιώντας ιεραρχική ομαδοποίηση σε παράλληλη μορφή. Όπως αναφέραμε στην παράγραφο 3.3.3, σε

Αλγόριθμος 8 Περιγραφή Παράλληλης Ιεραρχικής ομαδοποίησης

1: Σε κάθε επεξεργαστή ανατίθενται για υπολογισμό $(C * C)/P$ αποστάσεις όπου C ο αριθμός ομάδων και P ο αριθμός επεξεργαστών.
2: Ο κάθε επεξεργαστής υπολογίζει τις αποστάσεις που του έχουν ανατεθεί.
3: Ανακατασκευάζεται ο πίνακας αποστάσεων από τις αποστάσεις που έχει υπολογίσει ο κάθε επεξεργαστής.
4: Χρησιμοποιείται παράλληλος quick sort αλγόριθμος για να βρούμε τις ομάδες με τη μικρότερη απόσταση.
5: Ο κεντρικός επεξεργαστής αποφασίζει για τις νέες ομάδες και συνεχίζουμε στο βήμα 1 μέχρι να προκύψει μία μόνο ομάδα.

κάθε επανάληψη της ιεραρχικής ομαδοποίησης πρέπει να φτιάξουμε ένα πίνακα αποστάσεων μεταξύ των ομάδων. Στα πρώτα βήματα όπου οι ομάδες είναι πάρα πολλές, το υπολογιστικό κόστος εύρεσης του πίνακα αποστάσεων είναι υψηλό και μία παράλληλη επεξεργασία, επιταχύνει την διαδικασία. Η διαδικασία περιγράφεται από τον αλγόριθμο 8. Ο αλγόριθμος K-means έχει υλοποιηθεί σειριακά.

Αντιμετώπιση για πλήρη δεδομένα

Αν τα δεδομένα μας είναι πλήρη, ή έχουμε επιλέξει να τα ανακτήσουμε με K-NN τότε έχουμε τη δυνατότητα να επιλέξουμε μεταξύ συνεχών και διακριτών μεταβλητών. Στην περίπτωση των διακριτών μεταβλητών έχουμε μία επιλογή για την κατανομή της πυκνότητας πιθανότητας, την πολυωνυμική κατανομή. Στην περίπτωση των συνεχών μεταβλητών η κατανομή μας είναι μια υπό συνθήκη Gaussian κατανομή με γενική μορφή

$$p(x|PA_x) = c \cdot \exp(-\frac{1}{2\sigma^2}(x - \mu - f(PA_x))^2), \tag{7.1}$$

όπου c είναι μία σταθερά κανονικοποίησης και $f(PA_x)$ η συνάρτηση για τους πατέρες PA_x του x στο μοντέλο. Οι επιλογές για την f είναι πολλές. Οι βασικότερες και αυτές που χρησιμοποιήσαμε στο σύστημα μας είναι οι εξής:

1. Γραμμική ή πολυωνυμική,
2. Νευρωνικά δίκτυα [66],
3. B-Splines [22],
4. Radial Basis Kernel [67].

Αν η f είναι γραμμική, έχουμε την γραμμική Gaussian περίπτωση. Η εύρεση των παραμέτρων των πολυωνυμικών και γραμμικών Gaussian κατανομών για πλήρη δεδομένα έχει περιγραφεί στις παραγράφους 4.2.1 και 4.2.2. Για τις υπόλοιπες περιπτώσεις για να βρούμε τις παραμέτρους που μεγιστοποιούν την πιθανοφάνεια, χρησιμοποιήσαμε γνωστές μεθόδους

ελαχιστοποίησης όπως Gradient descent, BFGS, DFP. Όπως ήδη αναφέραμε η πιθανοφάνεια των δεδομένων D δοθέντος του μοντέλου M γράφεται ως εξής

$$\log p(D|M) = \log \int p(D|\Theta_M, M)p(\Theta_M|M)d\Theta_M.$$

Χρησιμοποιούμε την Laplace προσέγγιση του παραπάνω ολοκληρώματος στο σημείο $\widehat{\Theta}_M$ που μεγιστοποιείται η πιθανοφάνεια οπότε προκύπτει

$$\log p(D|M) = \sum_{i=1}^{m} \log p(D|\widehat{\Theta}_M, M) + \log p(\widehat{\Theta}_M|M) - \frac{d}{2}\log 2\pi - \frac{1}{2}log|A|, \qquad (7.2)$$

όπου $|A|$ η ορίζουσα του Εσσιανού (Hessian) πίνακα, d ο αριθμός των παραμέτρων στο μοντέλο.

Η διαδικασία που ακολουθείται είναι η εύρεση ενός τοπικού μεγίστου της $p(D|\widehat{\Theta}_M, M)$ και στην συνέχεια σε εκείνο το σημείο υπολογίζουμε τον Εσσιανό πίνακα. Αν αγνοήσουμε τους όρους που δεν εξαρτώνται από το αριθμό των δεδομένων N έχουμε στην ουσία το BIC μέτρο αξιολόγησης (ο όρος $-\frac{1}{2}\log|A|$ είναι ανάλογος του $-\frac{d}{2}\log N$).

Για το γραμμικό μοντέλο, τα νευρωνικά δίκτυα και τα B-Splines, χρησιμοποιήσαμε μέτρο BIC. Για τα Radial Basis όπου μπορούσαμε να υπολογίσουμε αναλυτικά την δεύτερη παράγωγο, χρησιμοποιήσαμε την προσέγγιση (7.2). Για τα νευρωνικά δίκτυα έχουμε στην ουσία μια εκ των προτέρων πιθανότητα στις παραμέτρους, να ανήκουν σε μια κανονική κατανομή με μέσο 0, και διασπορά 1. Για τα B-splines και Radial basis έχουμε μια εκ των προτέρων κατανομή εξομάλυνσης (smoothing) για τις παραμέτρους που περιγράφεται στην αναφορά [22], η οποία ευνοεί τις πιο ομαλές καμπύλες.

Αντιμετώπιση για ελλιπή δεδομένα

Στην περίπτωση που έχουμε ελλιπή δεδομένα, μπορούμε να τα ανακτήσουμε με την K-NN προσέγγιση είτε να χρησιμοποιήσουμε τον Structural EM αλγόριθμο. Ο αλγόριθμος μπορεί να διαχωριστεί στα ακόλουθα τμήματα:

- Εύρεση των ML ή MAP παραμέτρων.
- Αλγόριθμος αναζήτησης.
- Αξιολόγηση των υποψηφίων δικτύων
- Κατασκευή του συνδετικού δέντρου.
- Αλγόριθμος συμπερασματολογίας.

Η εύρεση των ML ή MAP παραμέτρων περιγράφεται στην παράγραφο 7.2. Ο αλγόριθμος αναζήτησης και αξιολόγησης των υποψηφίων δικτύων περιγράφεται στην παράγραφο 7.3. Τον αλγόριθμο συμπερασματολογίας τον περιγράψαμε στο Κεφάλαιο 6. Η κατασκευή του συνδετικού δέντρου είναι το μόνο κομμάτι του συστήματος που είναι σειριακό. Για ένα

Αλγόριθμος 9 Γενικός αλγόριθμος εύρεσης της καλύτερης δομής ενός Bayesian δικτύου

1: **repeat**
2: Υπολογίζουμε τις ML ή MAP παραμέτρους του τρέχοντος δικτύου
3: Βρίσκουμε τα εφικτά (που δεν σχηματίζουν κύκλο) νέα δίκτυα και κάθε επεξεργαστής αξιολογεί K/P υποψήφια δίκτυα όπου K ο συνολικός αριθμός των δικτύων και P ο αριθμός των επεξεργαστών.
4: Ο επεξεργαστής-συντονιστής λαμβάνει τα μέτρο αξιολόγησης για κάθε δίκτυο από όλους τους επεξεργαστές.
5: Ο επεξεργαστής-συντονιστής επιλέγει το καλύτερο υποψήφιο δίκτυο με μεγαλύτερο μέτρο αξιολόγησης από το τρέχον δίκτυο.
6: **until** Να μην υπάρχει υποψήφιο δίκτυο με μεγαλύτερο μέτρο αξιολόγησης

φυσιολογικό αριθμό κόμβων μέχρι περίπου 200, αυτό δεν αποτελεί πρόβλημα, αφού το υπολογιστικό κόστος της εύρεσης του συνδετικού δέντρου είναι πολύ μικρό σε σχέση με τις άλλες λειτουργίες. Κάθε επεξεργαστής έχει το δικό του συνδετικό δέντρο με συνεπή πάντα δυναμικά. Ο αλγόριθμος συμπερασματολογίας εκτελείται σε κάθε επεξεργαστή για τα δεδομένα που του έχουν ανατεθεί.

Τα βήματα που ακολουθούνται περιγράφονται συνοπτικά από τον αλγόριθμο 9 και σχηματικά στο Σχήμα 7.1.

7.2 Εύρεση ML ή MAP παραμέτρων

Για την εύρεση των ML ή MAP παραμέτρων χρησιμοποιείται παραλληλία στα δεδομένα. Δηλαδή «σπάμε» το αρχικό σύνολο δεδομένων σε P κομμάτια όσα είναι και οι επερξεργαστές. Σε αυτό το σημείο ο λόγος των δεδομένων προς τον αριθμό των επεξεργαστών $\frac{N}{P}$ θα πρέπει να είναι αρκετά μεγάλος ώστε να υπάρχει αυξημένο όφελος από την εκτέλεση σε παράλληλη μορφή. Αν ο λόγος είναι μικρός τότε η επιβάρυνση από την επικοινωνία μεταξύ των επεξεργαστών υπερκαλύπτει το όφελος της παραλληλίας. Στην παραλληλία εκμεταλλευόμαστε την ανεξαρτησία των δεδομένων σύμφωνα με την υπόθεση (4.4).

$$p(D|M) = \prod_{i=1}^{N} p(D_i|M), \tag{7.3}$$

$$\log p(D|M) = \sum_{i=1}^{N} \log p(D_i|M), \tag{7.4}$$

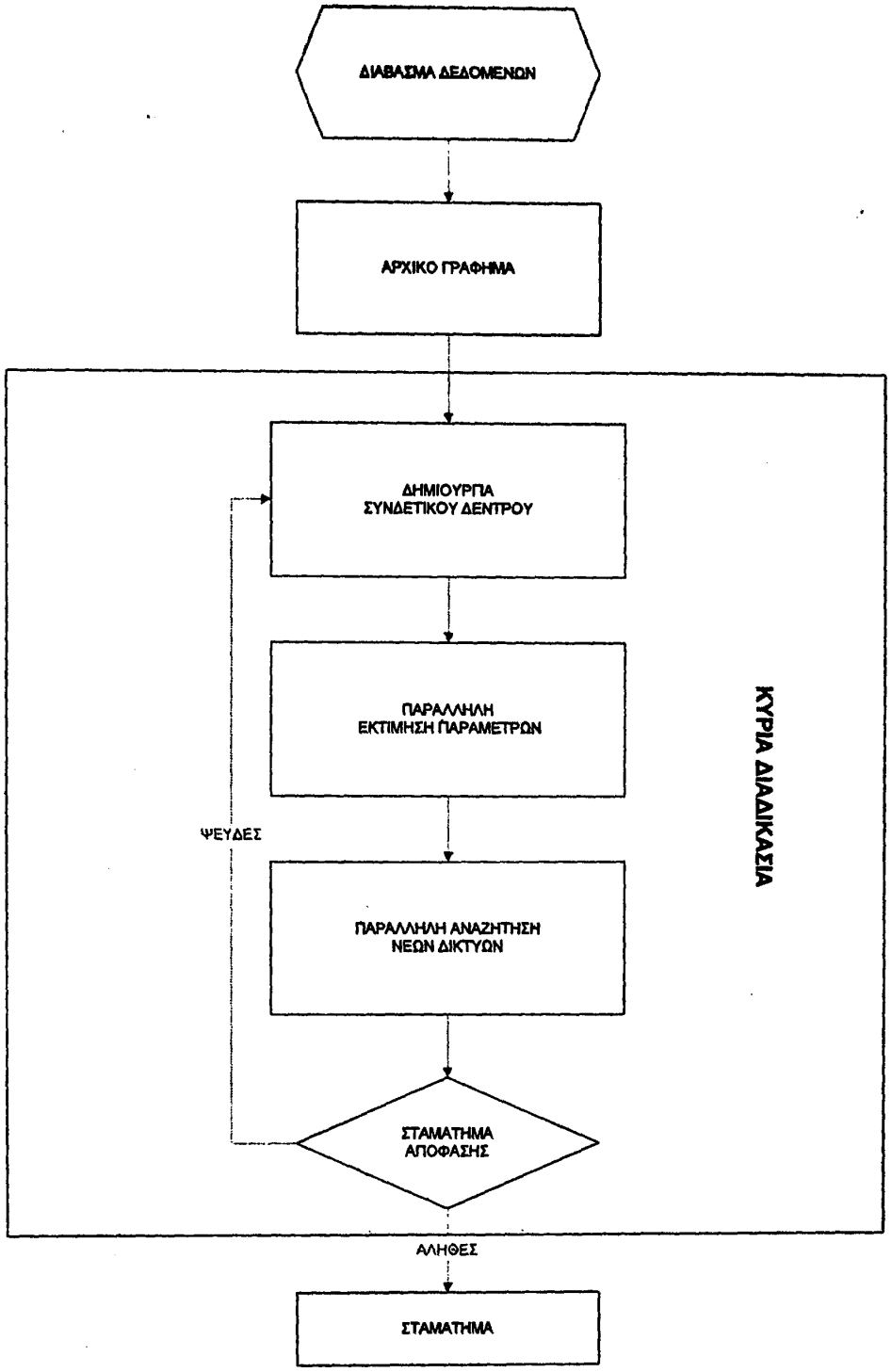
$$\log p(D|M) = \sum_{i=1}^{N/P} \log p(D_i|M) + \sum_{i=N/P}^{2*(N/P)} \log p(D_i|M), ..., \sum_{i=(N-1)*(N/P)}^{N} \log p(D_i|M) \tag{7.5}$$

Στην περίπτωση που έχουμε πλήρη δεδομένα και μη γραμμικά μοντέλα (π.χ. Radial Basis, B-Splines και ANN), χρησιμοποιούμε κάποιον αλγόριθμο βελτιστοποίησης, η παραλληλία των δεδομένων γίνεται στον υπολογισμό της συνάρτησης που ελαχιστοποιούμε και στον υπολογισμό της παραγώγου.

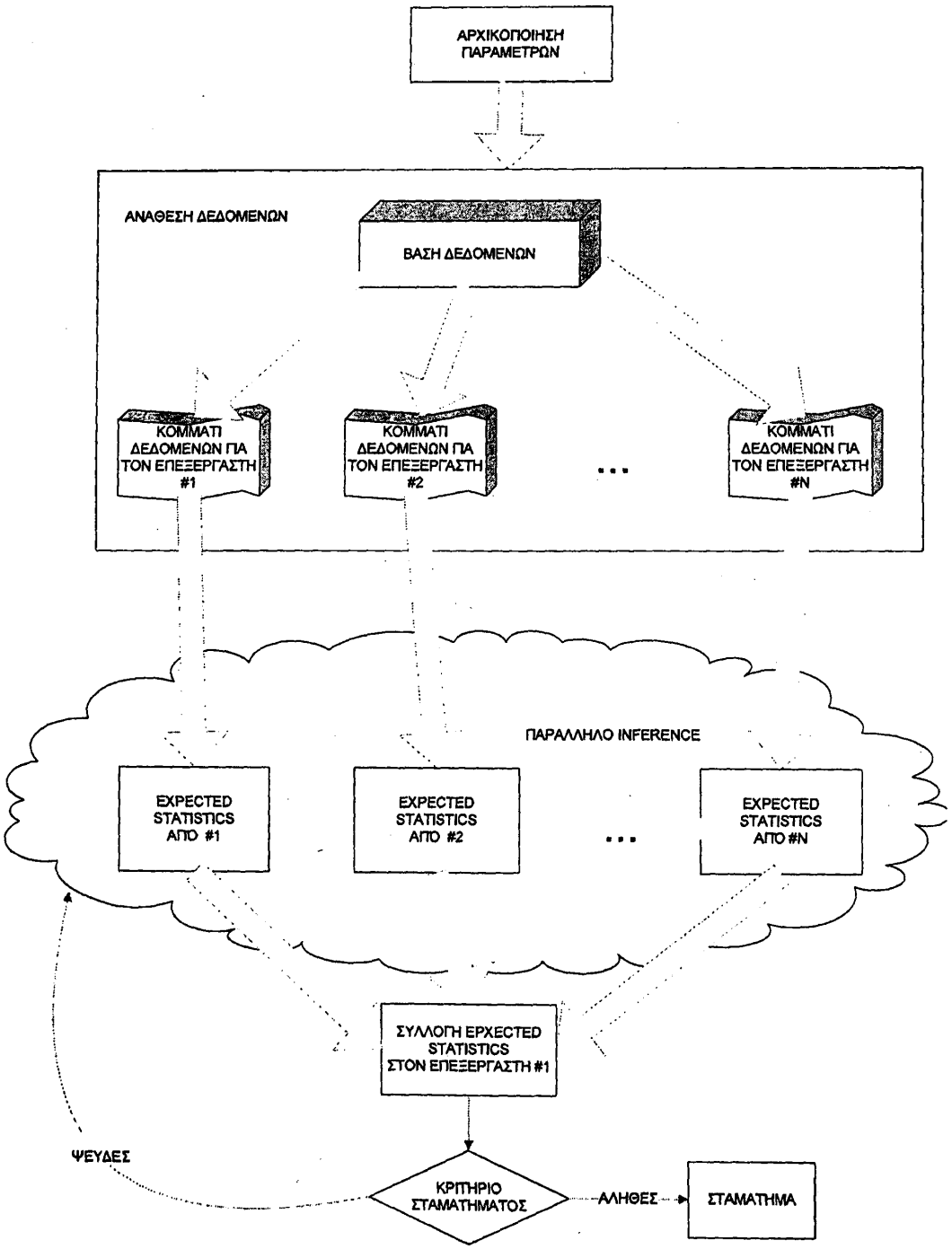
Στην περίπτωση που δεν έχουμε πλήρη δεδομένα κάθε επεξεργαστής έχει ένα αντίγραφο του συνδετικού δέντρου και υπολογίζει τα επαρκή στατιστικά για το σύνολο δεδομένων που του έχει ανατεθεί. Στην συνέχεια ο κεντρικός επεξεργαστής που εκτελεί και χρέη συντονιστή της διαδικασίας συλλέγει τα επιμέρους επαρκή στατιστικά και τα αθροίζει. Στη συνέχεια βρίσκει τις ML ή MAP παραμέτρους για κάθε κατανομή και ενημερώνει τους υπόλοιπους επεξεργαστές. Σχηματική περιγραφή φαίνεται στο Σχήμα 7.2.

7.3 Αναζήτηση και αξιολόγηση υποψηφίων δικτύων

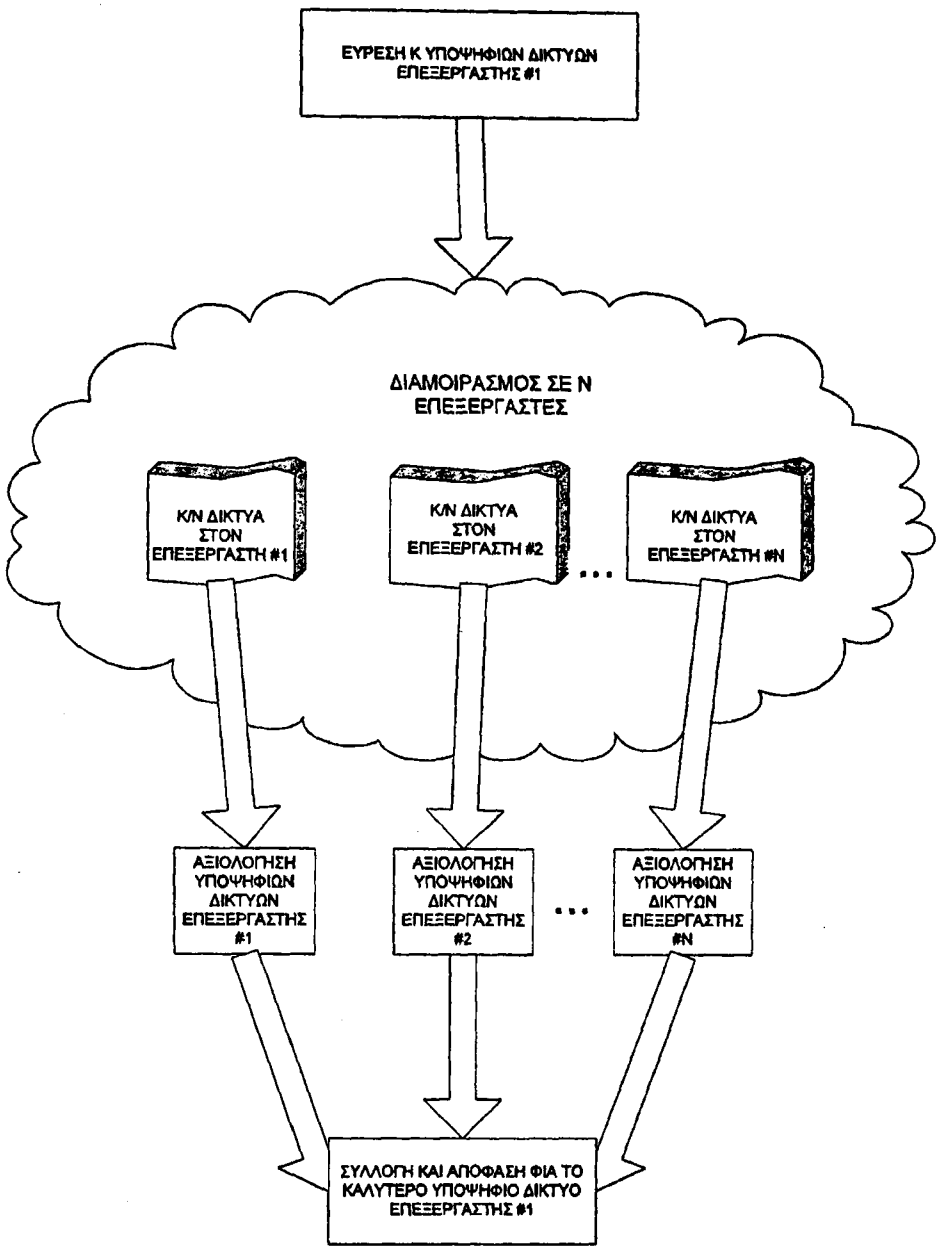
Αφού έχουν υπολογιστεί οι ML ή MAP παράμετροι για το τρέχον δίκτυο και έχουμε υπολογίσει το μέτρο αξιολόγησης κάθε μεταβλητής-κόμβου, στην συνέχεια πρέπει να αναζητήσουμε νέα δίκτυα και να επιλέξουμε το καλύτερο. Πρώτα ο κεντρικός επεξεργαστής αναλαμβάνει να αποφασίσει για υποψήφια δίκτυα που είναι επιτρεπτά (δεν δημιουργούν κύκλο). Εδώ πρέπει να σημειώσουμε ότι για πολύ μεγάλα δίκτυα, με περισσότερους από 100 κόμβους, υπάρχει η δυνατότητα (προς το παρόν για διακριτές μόνο κατανομές) να χρησιμοποιήσουμε μία εκδοχή του Sparse Candidate αλγορίθμου [2] για να μειώσουμε το υπολογιστικό κόστος. Έτσι αντί κάθε κόμβος να έχει N-1 υποψήφιους πατέρες, βρίσκουμε ένα υποσύνολο p μεταβλητών, όπου συνήθως $p \ll N-1$, βασιζόμενοι στο κριτήριο αμοιβαίας πληροφορίας (Mutual information criterion) [2]. Στην συνέχεια κάθε επεξεργαστής παίρνει ένα υποσύνολο υποψηφίων δικτύων και το αξιολογεί. Για την αξιολόγηση χρησιμοποιεί το αντίγραφο που έχει του συνδετικού δέντρου και των δυναμικών. Κάθε νέο δίκτυο διαφέρει στην γειτονιά ενός μόνου κόμβου-μεταβλητής, δηλαδή σε ένα μόνο τοπικό μοντέλο. Από το συνδετικό δέντρο και τα δυναμικά δημιουργείται η νέα κατανομή (6.3.3) και στην συνέχεια υπολογίζονται τα αναμενόμενα στατιστικά της νέας κατανομής. Η διαφορά με την εύρεση των αναμενόμενων στατιστικών στο βήμα όπου βρίσκουμε τις παραμέτρους του τρέχοντος δικτύου είναι ότι δεν εφαρμόζουμε σε κάθε πρότυπο τον αλγόριθμο συμπερασματολογίας αλλά όταν έχουμε ελλιπείς τιμές τις αντικαθιστούμε με τις αναμενόμενες τιμές που προκύπτουν όπως περιγράψαμε στην παράγραφο 6.3.3. Αφού έχουν βρεθεί τα αναμενόμενα επαρκή στατιστικά, υπολογίζονται οι παράμετροι της κατανομής και με τις παραμέτρους γνωστές μπορούμε να αξιολογήσουμε την νέα κατανομή. Η διαφορά στο μέτρο αξιολόγησης είναι το μέτρο αξιολόγησης της νέας κατανομής μείον αυτό της παλιάς του ίδιου κόμβου. Στην συνέχεια, ο κεντρικός επεξεργαστής συλλέγει τα μέτρα αξιολόγησης και αποφασίζει για το καλύτερο νέο δίκτυο, και ενημερώνει και τους υπόλοιπους για την απόφαση του. Σχηματικά η παραπάνω διαδικασία φαίνεται στο Σχήμα 7.3. Η επιλογή των υποψηφίων δικτύων γίνεται ανάλογα με τον αλγόριθμο που έχουμε επιλέξει. Για hill-climbing αλγόριθμο το σύνολο των νέων υποψηφίων δημιουργείται με προσθέσεις, αφαιρέσεις και εναλλαγές ακμών. Σημειώνουμε ότι σε κάθε επανάληψη, μετά την πρώτη, χρειάζεται να αξιολογήσουμε μόνο τα δίκτυα που προέρχονται από προσθέσεις, αφαιρέσεις και εναλλαγές ακμών των κόμβων που αλλάξαν κατά την προηγούμενη επανάληψη. Δηλαδή ενώ στο πρώτο βήμα, έχουμε $N \cdot (N-1)$ υποψήφια δίκτυα, στην συνέχεια έχουμε το πολύ $2 \cdot N$, γεγονός που διευκολύνει την διαδικασία μάθησης.



Σχήμα 7.1: Σχηματική περιγραφή αλγορίθμου 9.



Σχήμα 7.2: Παράλληλο σχήμα για την εύρεση των ML ή MAP παραμέτρων



Σχήμα 7.3: Παράλληλος ευρετικός αλγόριθμος

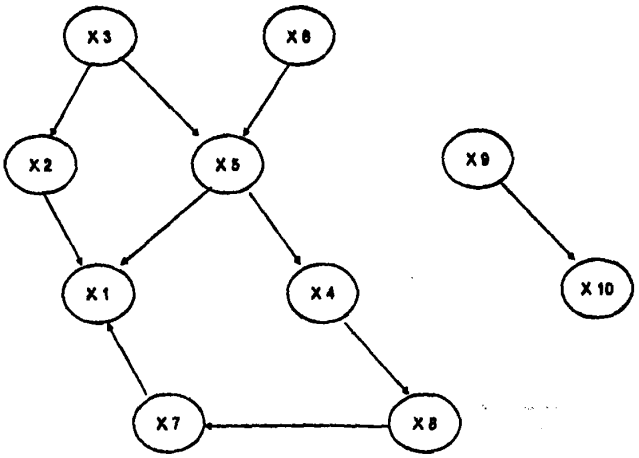
Κεφαλαιο 8

Πειραματα

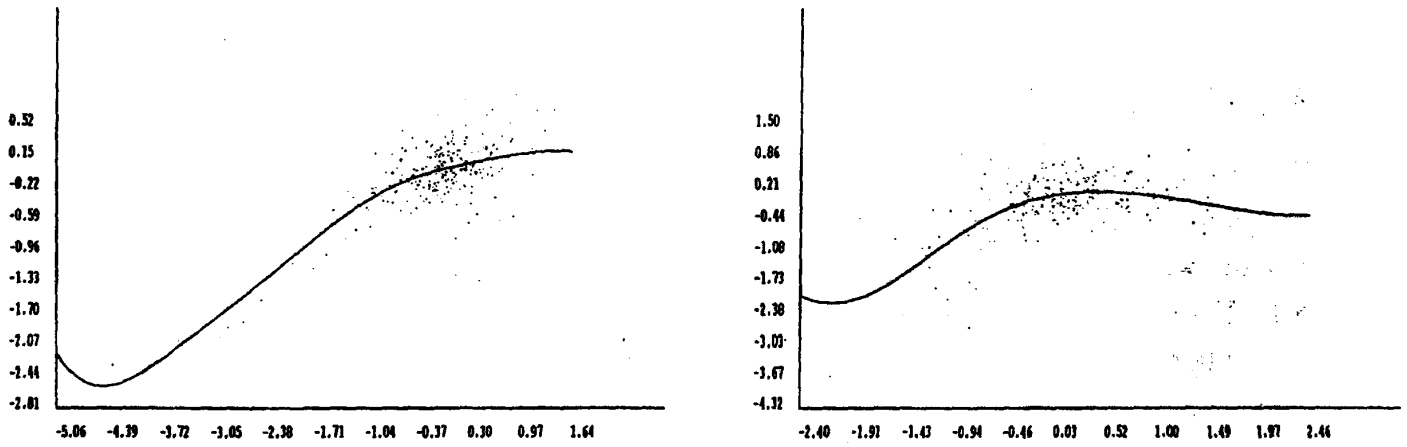
8.1 Πλήρη δεδομένα

Στο πρώτο μέρος των πειραμάτων συγκρίνουμε τις συναρτήσεις που χρησιμοποιήσαμε για την μοντελοποίηση της $f(Pa)$ στην εξίσωση (7.1) για μοντέλα συνεχών μεταβλητών. Για την σύγκριση αυτή χρησιμοποιήσαμε δεδομένα που πήραμε από το δίκτυο του Σχήματος 8.1. Στον Πίνακα 8.1 δίνονται οι κατανομές για κάθε κόμβο του δικτύου. Τρέξαμε τον αλγόριθμο εύρεσης του Bayesian δικτύου για κάθε κατανομή, με 100 επαναλήψεις στην κάθε μία. Επίσης θέσαμε $\sigma = 0.2$. Στον Πίνακα 8.1 παραθέτουμε τον μέσο αριθμό των ακμών που βρέθηκαν σωστά («TP») και τον μέσο αριθμό των ακμών που βρέθηκαν από τον αλγόριθμο ενώ δεν υπήρχαν στο πραγματικό δίκτυο («FP»). Επίσης φαίνεται και ο μέσος χρόνος που χρειάστηκε ο αλγόριθμος για κάθε μοντέλο.

Από τον Πίνακα 8.1 προκύπτει ότι όπως είναι αναμενόμενο τα νευρωνικά δίκτυα τα Radial Basis Kernels και τα B-Splines έχουν καλύτερα αποτελέσματα σε σχέση με γραμμικό μοντέλο αλλά είναι πιο αργά. Δεν προχωρήσαμε σε περισσότερα πειράματα γιατί τα



Σχήμα 8.1: Το δοκιμαστικό δίκτυο που χρησιμοποιήσαμε για να εξετάσουμε την απόδοση των συναρτήσεων που χρησιμοποιούμε για την μοντελοποίηση της κατανομήςχ πιθανότητας.



(α) (β)

Σχήμα 8.2: Δύο παραδείγματα αιτιατών σχέσεων μεταξύ γονιδίων. (α) YAL008W← FUN14 και (β) YAL035W← FUN12. Η φορά του βέλους από αριστερά προς τα δεξιά προσδιορίζει ότι το γονίδιο στα αριστερά του βέλους εξαρτάται από το γονίδιο στα δεξιά.

$$
\begin{aligned}
X_1 &= X_2^2 + 2\sin(X_5) - 2X_7 + N(0,\sigma^2) \\
X_2 &= \{1+\exp(-4X_3)\}^{-1} + N(0,\sigma^2) \\
X_3 &= N(0,\sigma^2) \\
X_4 &= X_5^2/3 + N(0,\sigma^2) \\
X_5 &= X_3 - X_6^2 + N(0,\sigma^2) \\
X_6 &= N(0,\sigma^2) \\
X_7 &= \begin{cases} -1 + N(0,\sigma^2) & (X_8 \le -0.5) \\ X_8 + N(0,\sigma^2) & (-0.5 \le X_8 \le 0.5) \\ 1 + N(0,\sigma^2), & (\le 0.5 X_8) \end{cases} \\
X_8 &= \exp(-X_4 - 1)/2 + N(0,\sigma^2) \\
X_9 &= N(0,\sigma^2) \\
X_{10} &= \cos(X_9) + N(0,\sigma^2)
\end{aligned}
\tag{8.1}
$$

Πίνακας 8.1: Οι συναρτήσεις από τις οποίες προέρχονται τα δεδομένα για κάθε κόμβο του δικτύου 8.1. Η συνάρτηση που παράγει δεδομένα για κάθε μεταβλητή-κόμβο προκύπτει από συνδυασμό συναρτήσεων για κάθε πατέρα του κόμβου στο γράφημα με την πρόσθεση θορύβου μέσου 0 και διασποράς σ.

Μοντέλο $f(Pa)$	TP	FP	ΧΡΟΝΟΣ
Linear	6.22	3.52	4.89
ANN	8.81	2.25	278.12
B-Splines	8.63	2.58	60.78
Radial Basis Kernels	9.18	2.10	58.34

Πίνακας 8.2: Σύγκριση μεθόδων στο δίκτυο 8.1 για 100 επαναλήψεις με $\sigma = 0.2$ και 1000 δεδομένα. Με «TP» δηλώνεται το ποσοστό των ακμών που βρέθηκαν σωστά από την κάθε μέθοδο, με «FP» ο αριθμός των ακμών του δικτύου που δεν υπάρχουν στο αρχικό δίκτυο και ο «χρόνος» είναι ο μέσος χρόνος που χρειάστηκε ο αλγόριθμος για το αντίστοιχο μοντέλο.

Μοντέλο $f(Pa)$	Υπολογιστικό κόστος	Ακρίβεια
Linear	✓	
ANN		✓
B-Splines	✓	✓
Radial Basis Kernels	✓	✓

Πίνακας 8.3: Σύγκριση πλεονεκτήματων για τα μοντέλα που χρησιμοποιούμε για την $f(Pa_X)$.

αποτελέσματα είναι αναμενόμενα και τα μοντέλα που χρησιμοποιήσαμε έχουν μελετηθεί εκτενώς στην βιβλιογραφία. Εδώ πρέπει να σημειώσουμε ότι στο πείραμα μας τα Radial Basis Kernels έχουν καλύτερη απόδοση γιατί χρησιμοποιήσαμε το μέτρο αξιολόγησης που δίνει η Σχέση (7.2) και που περιέχει περισσότερη πληροφορία από το BIC μέτρο. Γενικά οι αιτιατές σχέσεις μεταξύ γονιδίων όπως φαίνεται και από τα Σχήματα 8.2(α)-(β) δεν είναι πολύπλοκες, και μπορούν να καθοριστούν από απλά γραμμικά μοντέλα. Έτσι η χρήση πολύπλοκων μοντέλων, όπως τα νευρωνικά δίκτυα, που είναι και πιο απαιτητικά από άποψη υπολογιστικού χρόνου, δεν είναι υπολογιστικά αποδοτική. Μια καλή λύση είναι η χρήση B-Splines ή Radial Basis Kernels και ειδικά τα Radial Basis Kernels όπου μπορούμε να υπολογίσουμε αναλυτικά τον Hessian πίνακα.

8.2 Σύγκριση του SEM με την κλασσική προσέγγιση για ελλιπείς τιμές

Όταν τα δεδομένα μας δεν είναι πλήρη, τότε μπορούμε να ακολουθήσουμε δύο προσεγγίσεις:

1. Εφαρμόζουμε κάποια από τις μεθόδους που αναφέραμε για ανάκτηση των χαμένων τιμών (KNN, SVD) και στην συνέχεια στα πλήρη δεδομένα ψάχνουμε να βρούμε το «καλύτερο» Bayesian δίκτυο

2. Χρησιμοποιούμε τον Structural EM για την μάθηση Bayesian δικτύων με μη πλήρη δεδομένα.

Το πλεονέκτημα της πρώτης μεθόδου είναι ότι εκμεταλλεύεται πλήρως όλες τις ιδιότητες των Bayesian δικτύων, γεγονός που μας επιτρέπει την χρήση των γρήγορων αλγορίθμων αναζήτησης και αξιολόγησης δικτύων. Γενικά όταν έχουμε μικρό ποσοστό χαμένων τιμών οι μέθοδοι αποκατάστασης χαμένων τιμών δίνουν πολύ καλά αποτελέσματα οπότε η χρήση αυτής της προσέγγισης ενδείκνυται.

Από την άλλη όταν χειριζόμαστε τις ελλιπείς τιμές μέσα στην διαδικασία της εύρεσης ενός ή μερικών "καλών" δικτύων που μεγιστοποιούν την πιθανοφάνεια των δεδομένων, έχουμε το πλεονέκτημα ότι για κάθε μεταβλητή έχουμε και ένα σύνολο πατέρων από το οποίο εξαρτάται και κατά συνέπεια υπάρχει και καλύτερη εξήγηση για την κατανομή που δημιουργεί τα δεδομένα σε σχέση με τις μεθόδους ανάκτησης χαμένων τιμών. Επιπλέον

έχει αποδειχθεί ότι οι ελλιπείς τιμές πολλές φορές βοηθούν στην αποφυγή υπερεκπαίδευσης. Το βασικό μειονέκτημα είναι ότι η μάθηση με ελλιπείς τιμές είναι λίγο ως πολύ πιο αργές από τη μάθηση με πλήρες σύνολο δεδομένων. Εμείς δοκιμάσαμε και συγκρίναμε και τις δύο κατηγορίες μεθόδων. Σε αυτό το σύνολο πειραμάτων θα προσπαθήσουμε να βγάλουμε κάποια συμπεράσματα για την απόδοση του SEM. Πρώτα θα συγκρίνουμε το SEM όταν έχουμε συνεχείς μεταβλητές και συνεπώς γραμμικά Gaussian μοντέλα με την κλασική προσέγγιση όπου ανακτούμε πρώτα τις ελλιπείς τιμές με KNN. Για να αξιολογήσουμε αυτές τις δύο μεθόδους φτιάξαμε ένα τυχαίο δίκτυο με 20 κόμβους που έχει περίπου την μορφή δέντρου και έχει συνολικά 20 ακμές. Ο μέγιστος αριθμός πατέρων για κάθε μεταβλητή είναι 2, οπότε το δίκτυο μας είναι σχετικά απλό, και οι απαιτήσεις για το σύνολο εκπαίδευσης είναι μικρές. Αυτό φαίνεται στα πειράματα. Στο Σχήμα 8.3 βλέπουμε την πιθανοφάνεια στο σύνολο test για 250, 500, 1000, 2000 δείγματα συνόλου εκπαίδευσης, τόσο για την πρώτη περίπτωση όσο και για την δεύτερη. Από το σχήμα βγαίνουν τρία πρώτα συμπεράσματα

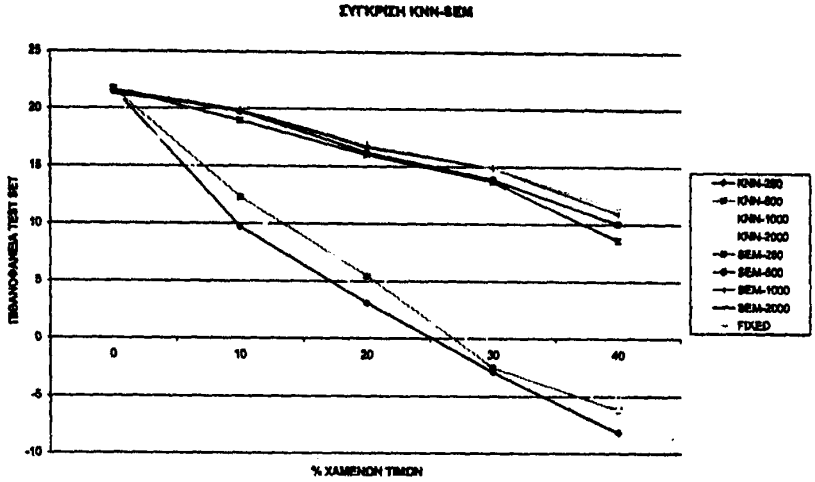
1. Η απόδοση του SEM είναι μακράν καλύτερη από αυτής της ανάκτησης με KNN. Αυτό είναι γενικά αναμενόμενο αφού ο SEM στην ουσία αναπληρώνει τις ελλιπείς τιμές με αυτές που μεγιστοποιούν την πιθανοφάνεια του μοντέλου της κάθε επανάληψης του αλγορίθμου. Από την άλλη η αναμενόμενη διαφορά στην απόδοση θα ήταν μικρότερη αν ο αριθμός των μεταβλητών ήταν μεγαλύτερος συγκριτικά με τον όγκο των δεδομένων οπότε ο KNN θα μπορούσε να αποδώσει καλύτερα.

2. Η δεύτερη παρατήρηση είναι ότι επειδή το δίκτυο είναι αρκετά απλό η διαφορά στην απόδοση σε σχέση με τον αριθμό των δειγμάτων στο σύνολο εκπαίδευσης είναι μικρή, και αυξάνει λίγο όταν έχουμε πολλές ελλιπείς τιμές.

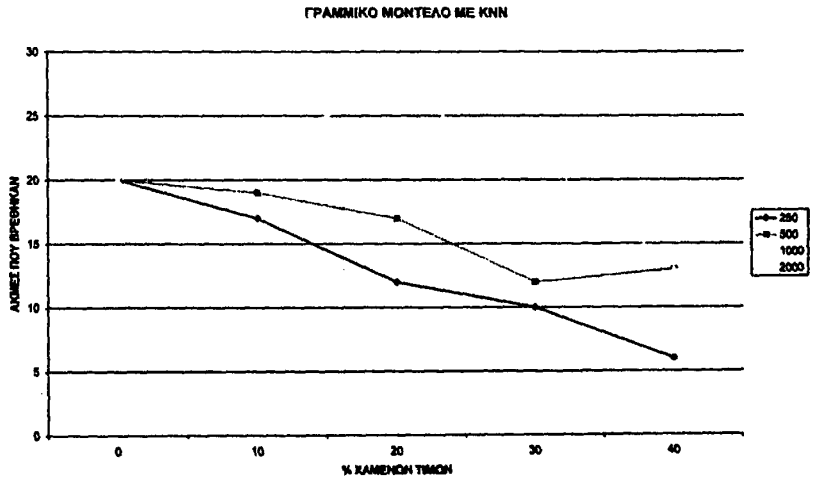
3. Η καμπύλη με ένδειξη FIXED είναι η μεγαλύτερη πιθανοφάνεια στο σύνολο δοκιμής (test set) (αντιστοιχεί στα 2000 δείγματα του συνόλου εκπαίδευσης) για το δίκτυο με πραγματική δομή, αφού βρήκαμε τις ML παραμέτρους του. Αυτό σημαίνει ότι είμαστε πολύ κοντά στην καλύτερη απόδοση που θα μπορούσε να έχει ο συγκεκριμένος αλγόριθμος και με αυτό το σύνολο δεδομένων.

8.3 SEM και διακριτές μεταβλητές

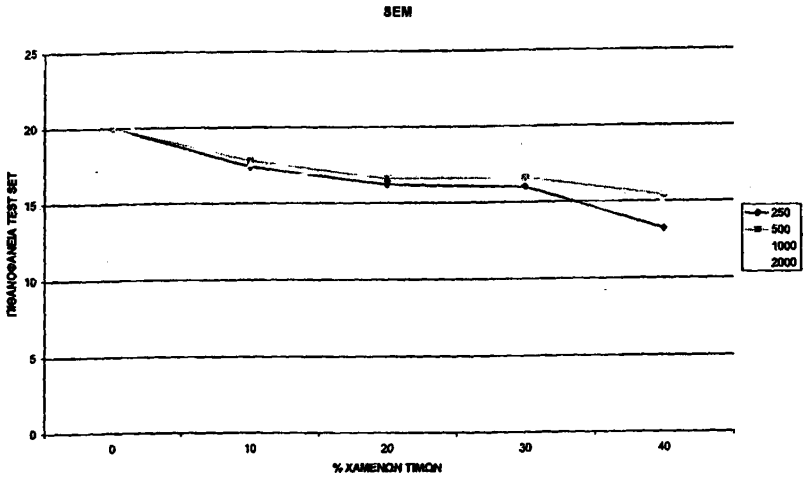
Παρακάτω παραθέτουμε πειράματα για την αξιολόγηση της απόδοσης του SEM για διακριτές (πολυωνυμικές) κατανομές. Για να ελέγξουμε την απόδοση του αλγορίθμου τον δοκιμάσαμε σε τρία γνωστά δίκτυα (ALARM, ASIA, CAR DIAGNOSIS) και ελέγξαμε την πιθανοφάνεια του εκπαιδευμένου δικτύου και τον αριθμό των σωστών ακμών που βρήκε



Σχήμα 8.3: Σύγκριση της πιθανοφάνειας στο σύνολο test των δικτύων που βρίσκει ο SEM με αυτά του γραμμικού μοντέλου αφού έχουμε ανακτήσει τις ελλιπείς τιμές με KNN (σε σχέση με το ποσοστό των χαμένων τιμών στο σύνολο εκπαίδευσης).



Σχήμα 8.4: Οι ακμές που βρίσκει το γραμμικό μοντέλο με KNN

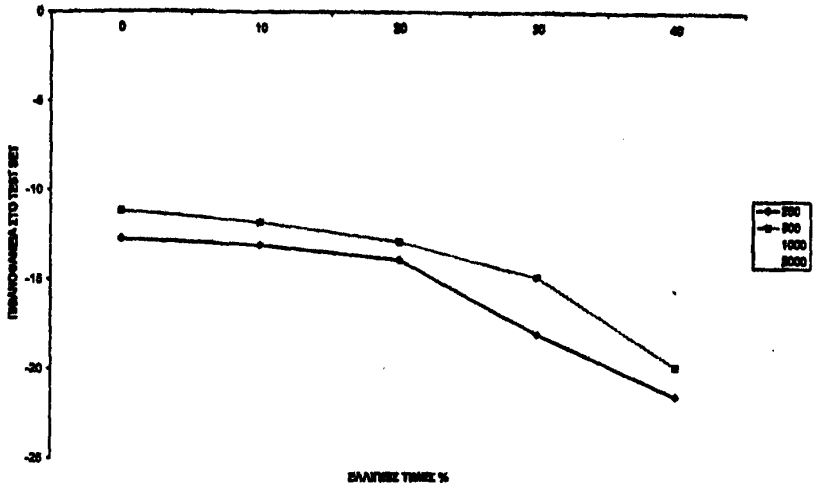


Σχήμα 8.5: Οι ακμές που βρίσκει ο SEM

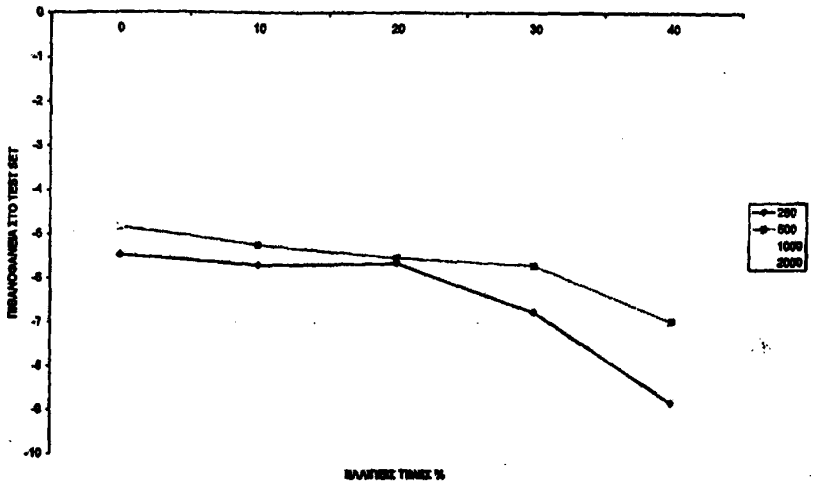
ΔΙΚΤΥΟ	ΚΟΜΒΟΙ	ΑΚΜΕΣ
ALARM	37	46
CAR DIAGNOSIS	17	18
ASIA	8	8

Πίνακας 8.4: Δίκτυα διακριτών μεταβλητών για αξιολόγηση του SEM

ο αλγόριθμος σε συνάρτηση με το ποσοστό των χαμένων τιμών στο σύνολο εκπαίδευσης. Τα τρία δίκτυα που χρησιμοποιήσαμε δίνονται στον Πίνακα 8.3. Τα αποτελέσματα που προκύπτουν είναι τα αναμενόμενα. Όσο περισσότερα δεδομένα έχουμε και όσο μικρότερο είναι το ποσοστό των χαμένων τιμών τόσο καλύτερη είναι και η απόδοση του SEM. Τα αποτελέσματα είναι παρόμοια με αυτά του Friedman [6]. Όταν τα δεδομένα είναι σχετικά επαρκή σε σχέση με την πολυπλοκότητα του δικτύου παρουσιάζεται μια σχετική σταθερότητα τόσο της πιθανοφάνειας του δικτύου όσο και του αριθμού των ακμών που βρίσκονται σωστά, σε σχέση με το ποσοστό των ελλιπών τιμών, όσο αυτό το ποσοστό φυσικά κυμαίνεται σε ένα φυσιολογικό πλαίσιο μέχρι περίπου 30%.



Σχήμα 8.6: Η πιθανοφάνεια για το ALARM δίκτυο ως συνάρτηση του ποσοστού των χαμένων τιμών



Σχήμα 8.7: Η πιθανοφάνεια για το CAR DIAGNOSIS δίκτυο ως συνάρτηση του ποσοστού των χαμένων τιμών

Alarm				
ΧΑΜ.ΤΙΜΕΣ %\ΔΕΔΟΜΕΝΑ	250	500	1000	2000
0%	16.3	20.7	23.1	25.2
10%	16.3	20.7	23.1	25.2
20%	12.3	14.6	17.3	19.3
30%	5.5	10.5	13.4	14.8
40%	0.4	3.1	7	10.5

(α)

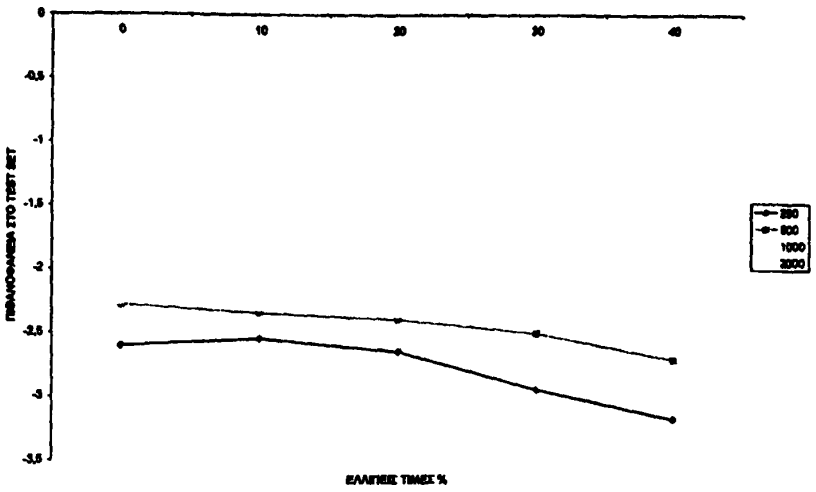
CAR DIAGNOSIS				
ΧΑΜ.ΤΙΜΕΣ % \ΔΕΔΟΜΕΝΑ	250	500	1000	2000
0 %	8	11	13	13
10 %	5.8	7.5	9	10.3
20%	6	6	7.6	8.6
30%	4.2	5.5	5.6	6.9
40%	1.6	3.8	4.4	5.3

(β)

ASIA				
ΧΑΜ.ΤΙΜΕΣ % \ΔΕΔΟΜΕΝΑ	250	500	1000	2000
0%	4	6	7	7
10%	3.1	4.1	5.6	5.7
20%	2.6	3.6	4.7	5.4
30%	1.2	2.7	3.9	4.8
40%	0.7	1.6	2.2	3.2

(γ)

Πίνακας 8.5: Αριθμός ακμών που βρέθηκαν σωστά ως συνάρτηση του ποσοστού των χαμένων τιμών και του αριθμού των δεδομένων. (α) Για το Alarm δίκτυο, (β) για το CAR DIAGNOSIS δίκτυο,(γ) για το ASIA το δίκτυο.

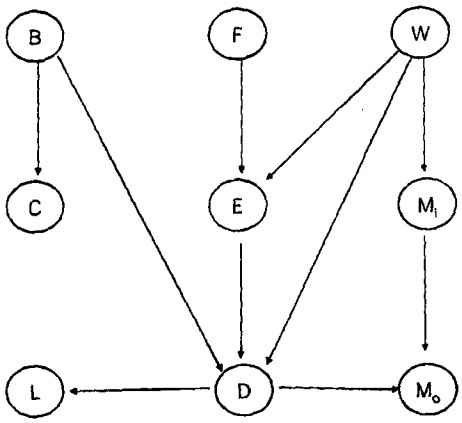


Σχήμα 8.8: Η πιθανοφάνεια για το ASIA δίκτυο ως συνάρτηση του ποσοστού των χαμένων τιμών

8.4 Υβριδική περίπτωση

Σε αυτό το μέρος των πειραμάτων, εξετάζουμε τον υβριδικό αλγόριθμο και την απόδοση του σε τεχνητά δίκτυα. Το τεχνητό δίκτυο που δοκιμάσαμε, είναι αυτό του Σχήματος 8.9, που χρησιμοποιήθηκε από τον Lauritzen [61]. Ήδη έχουμε περιγράψει στο κεφάλαιο 6.4 τον υβριδικό αλγόριθμο. Η πρώτη παρατήρηση που έχουμε να κάνουμε από τα πειράματα είναι ότι για την εκπαίδευση ενός τέτοιου δικτύου χρειάζονται περισσότερα δεδομένα από τις προηγούμενες περιπτώσεις. Ο λόγος είναι ότι στην περίπτωση που μια συνεχής μεταβλητή έχει ένα ή περισσότερους διακριτούς πατέρες τότε προκύπτει μια CG κατανομή. Τότε σε κάθε πιθανό σύνολο τιμών των πατέρων έχουμε και μια Gaussian κατανομή της οποίας τις παραμέτρους πρέπει να βρούμε. Αν κάποιο τέτοιο σύνολο τιμών δεν εμφανίζεται επαρκώς στα δεδομένα δε θα μπορέσουμε να βρούμε αποτελεσματικά τις παραμέτρους της gaussian κατανομής και αυτό είναι ένα πολύ συχνό φαινόμενο. Από την άλλη θα μπορούσε κάποιος να υποστηρίξει ότι από την στιγμή που ένα σύνολο πιθανών τιμών των διακριτών πατέρων έχει μικρή πιθανότητα, δεν μας ενδιαφέρει να βρούμε την κατανομή των συνεχών μεταβλητών σε αυτή την περίπτωση.

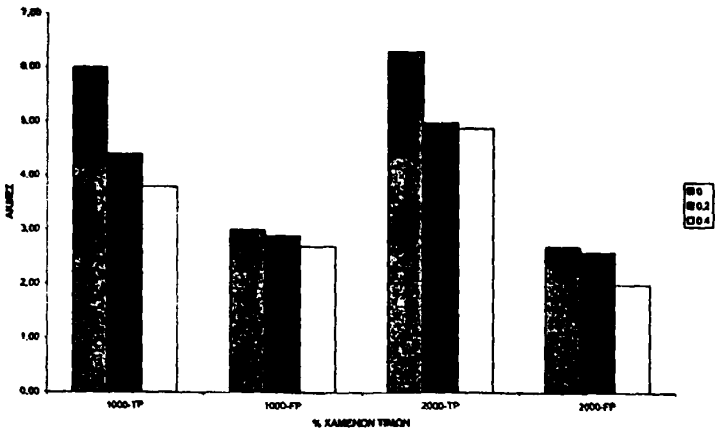
Ας προχωρήσουμε στα συμπεράσματα από τα πειράματα. Δοκιμάσαμε τον αλγόριθμο στο δίκτυο του Σχήματος 8.9 για 1000 και 2000 δεδομένα. Στο Σχήμα 8.10(α) βλέπουμε τις ακμές που βρέθηκαν σωστά και τις ακμές που εμφανίζονται, αλλά δεν υπάρχουν στο αρχικό δίκτυο σε σχέση με το ποσοστό των χαμένων τιμών. Ο αριθμός των ακμών του πραγματικού δικτύου είναι 10. Ο αλγόριθμος βρίσκει μέχρι 6.32 (μέσος όρος) ακμές με την σωστή κατεύθυνση. Στις ακμές που εμφανίζονται αλλά δεν αντιστοιχούν σε πραγματικές, περιέχονται και αυτές που έχουν λάθος κατεύθυνση. Γενικά η απόδοση του αλγορίθμου δεν



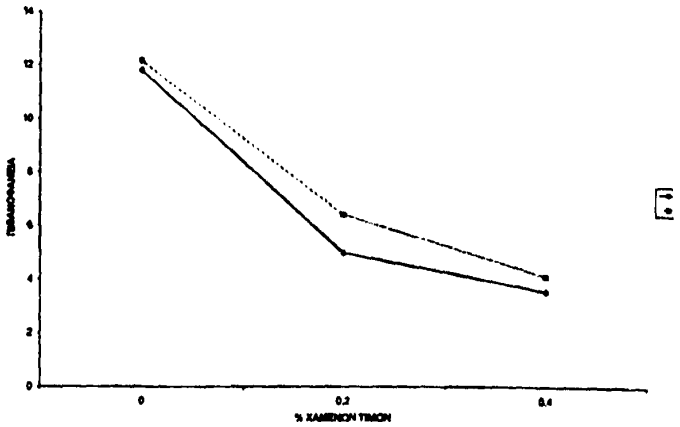
Σχήμα 8.9: Το διακριτό δίκτυο που χρησιμοποιήσαμε στα πειράματα [61]. Οι μεταβλητές είναι W (τύπος αποβλήτου), F (κατάσταση φίλτρου), B (κάυση), M_i (μέταλλα στα απόβλητα), E (αποτελεσματικότητα του φίλτρου), C (εκπομπή $C0_2$), D (εκπομπή σκόνης), M_0 (εκπόμπή μετάλλων) και L (διαπεραστικότητα από φως). Οι μεταβλητές W,F και B είναι διακριτές, οι υπόλοιπες συνεχείς.

έχει μεγάλη διαφορά με αυτήν του αλγορίθμου για μόνο συνεχείς μεταβλητές. Στο Σχήμα 8.10(β) έχουμε την πιθανοφάνεια (για το test σύνολο) του δικτύου σε σχέση με το ποσοστό χαμένων τιμών για 1000 και 2000 δεδομένα.

Στόχος μας είναι η εύρεση των γενετικών δικτύων. Η αρχική ιδέα για την χρήση των υβριδικών δικτύων ήταν να τις παραστήσουμε ως διακριτές μεταβλητές πότε μια μεταβλητή λείπει ή όχι. Με αυτόν τον τρόπο θα μπορούσαμε να εξακριβώσουμε αν ισχύει η υπόθεση ότι τα χαμένα δεδομένα στο πρόβλημα παράγονται τυχαία. Η διαίσθηση λέει ότι κάτι τέτοιο ισχύει εν μέρει και συγκεκριμένα θα μπορούσε να δικαιολογηθεί από το γεγονός ότι η τεχνολογία των μικροσυστοιχιών DNA δεν εγγυάται ότι δεν υπάρχει θόρυβος στις εικόνες τέτοιος ώστε να υπάρξει απώλεια κάποιων τιμών. Είναι γεγονός όμως ότι είναι πιο πιθανό κάποια δεδομένα να λείπουν επειδή το αντίστοιχο γονίδιο δεν εκφράζεται. Οπότε μια τέτοια διαδικασία δεν μπορούμε να την εκλάβουμε σαν τυχαία. Επειδή όμως τα δεδομένα από τις μικροσυστοιχίες DNA δεν είναι επαρκή, προκύπτουν αριθμητικά λάθη στην εφαρμογή του αλγορίθμου. Θα μπορέσουμε να αξιολογήσουμε την υπόθεση όταν συλλέξουμε έναν ικανοποιητικό αριθμό δεδομένων. Ένα ακόμα μειονέκτημα αυτής της μεθόδου είναι ότι δεν επιτρέπει σε διακριτές μεταβλητές να έχουν συνεχείς πατέρες. Θα πρέπει να επεκτείνουμε την μέθοδο ώστε να άρουμε αυτόν τον περιορισμό [68].



(α)

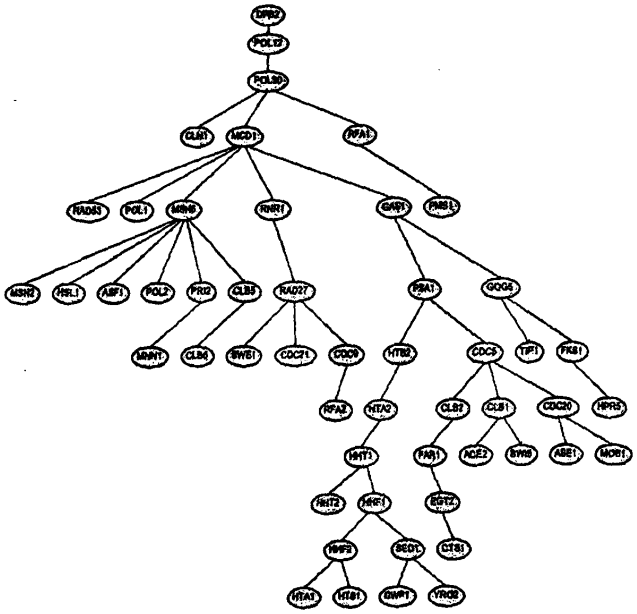


(β)

Σχήμα 8.10: (α)Οι ακμές που βρέθηκαν σωστά και οι ακμές που βρέθηκαν, ενώ δεν υπάρχουν στο πραγματικό δίκτυο, για 1000 και 2000 δεδομένα. και (β)Η πιθανοφάνεια του test συνόλου, για 1000 και 2000 δεδομένα σε σχέση με τον ποσοστό χαμένων δεδομένων .

8.5 Δίκτυα Γονιδίων

Εφαρμόζουμε τις μεθόδους μας σε δεδομένα που έχουμε από μικροσυστοιχίες DNA και συγκεκριμένα από τον κύκλο του κυττάρου στη ζύμη (*S. cerevisiae*). Συγκεκριμένα τα δεδομένα έχουν συλλεχθεί από τρεις βάσεις ([3], [4], [5]), αλλά δεν αφορούν όλες τον κύκλο του κυττάρου. Τα γνωστά γονίδια που έχουν κάποιο ρόλο στον κύκλο του κυττάρου είναι 108. Και τα 108 περιέχονται στην [3]. Οπότε έχουμε προχωρήσει στα παρακάτω πειράματα. Τα κοινά και στις τρεις βάσεις είναι 88. Δοκιμάσαμε με το σύνολο δεδομένων και από τις τρεις και και μόνο από αυτήν με τον κύκλο του κυττάρου και θα προσπαθήσουμε να δούμε πόσο ταιριάζουν τα αποτελέσματα μας με την βιβλιογραφία. Στα Σχήματα 8.11, 8.12, 8.13, δίνουμε τμήματα των δικτύων που πήραμε από τα δεδομένα με διακριτές, γραμμικές και Radial Basis κατανομές αντίστοιχα. Στα Σχήματα 8.14(α)-(β), 8.15 παραθέτουμε την γειτονιά του γονιδίου CLN1 που είναι γνωστό ότι ενεργοποιείται στην αρχή του κύκλου του κυττάρου (πολυωνυμικά, γραμμικά και Radial Basis μοντέλα) και στα Σχήματα 8.16(α)-(β), 8.17 παραθέτουμε την γειτονιά του γονιδίου HTA1, αντίστοιχα. Παρατηρούμε ότι οι βασικές σχέσεις εμφανίζονται και με τις τρεις κατανομές που αναφέραμε παραπάνω. Όπως και με το αρχικό δίκτυο μπορούμε να παρατηρήσουμε ότι με συνεχείς μεταβλητές το δίκτυο μας γίνεται πιο συνεκτικό και κάθε γονίδιο εμφανίζεται με μεγαλύτερη γειτονιά σε σχέση με τις αντίστοιχες με Radial Basis και διακριτές πολυωνυμικές κατανομές. Μπορούμε να παρατηρήσουμε ότι στην γειτονιά τόσο του CLN1 όσο και του HTA1 εμφανίζονται γονίδια γειτονικά στο χρωμόσωμα. Στο Σχήμα 8.18 παραθέτουμε το δίκτυο για τα 800 γονίδια που εκφράζονται στον κύκλο του κυττάρου σύμφωνα με την [20]. Στον Πίνακα 8.5 παραθέτουμε τα σκορ για τα 88 γονίδια του δικτύου που βρήκε ο αλγόριθμος με διακριτές (πολυωνυμικές)

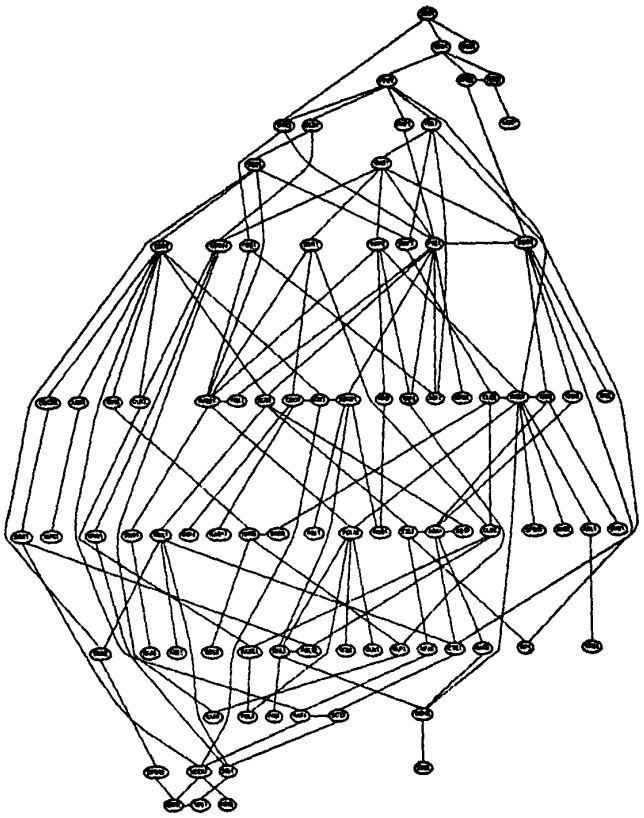


Σχήμα 8.11: Ένα μέρος του γενετικού δικτύου του κύκλου του κυττάρου που πήραμε χρησιμοποιώντας πολυνυμικές κατανομές.

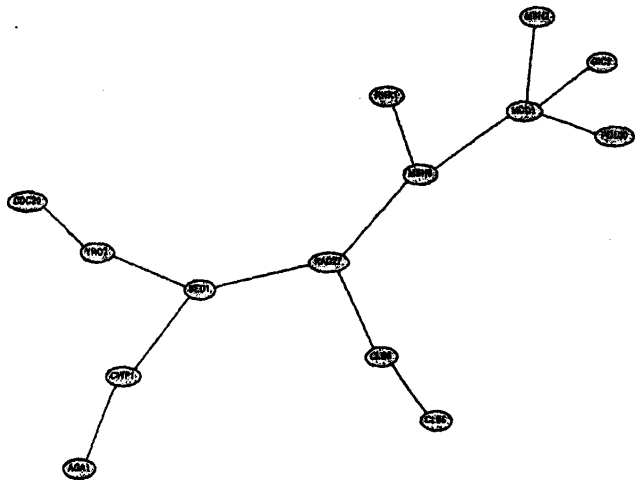
κατανομές.

8.6 Απόδοση παραλληλίας

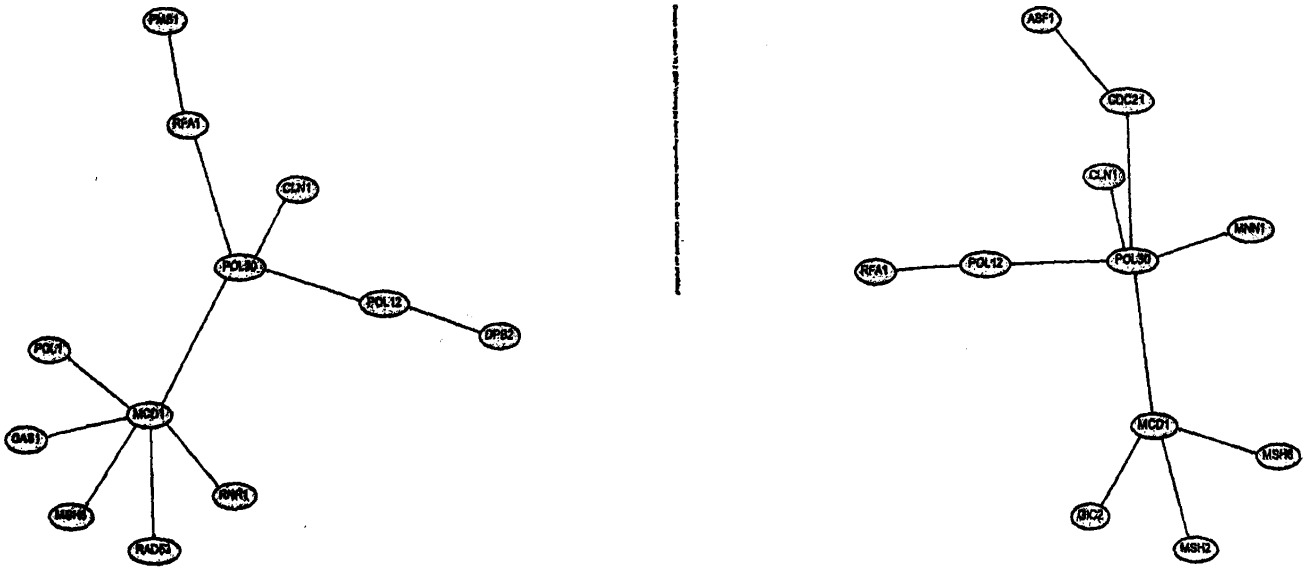
Σε αυτό το κεφάλαιο παραθέτουμε κάποιες πληροφορίες όσο αφορά την απόδοση του συστήματος μας και ειδικότερα την απόδοση της παραλληλίας. Ήδη έχουμε αναφερθεί στα προβλήματα που παρουσιάζονται και στην παράγραφο 8.7 όπου θέτουμε και τους στόχους για την μελλοντική εργασία, εξετάζουμε θεωρητικά αν και με ποιο τρόπο αυτά θα μπορούσαν να λυθούν. Παρακάτω παρουσιάζουμε διαγράμματα με τον χρόνο εκτέλεσης του συνολικού αλγορίθμου αλλά και για διάφορα υπό-τμήματα του συνολικού συστήματος, σε σχέση με διάφορους παράγοντες. Οι παράγοντες που εξετάσαμε και που παίζουν σημαντικό ρόλο στην απόδοση του αλγορίθμου είναι ο αριθμός των δεδομένων (500, 1000, 2000), ο αριθμός των κόμβων (20, 40, 80) και το ποσοστό των χαμένων τιμών (0%, 20%, 40%). Για την αξιολόγηση του αλγορίθμου όσο αφορά τον αριθμό των δεδομένων και το ποσοστό των χαμένων τιμών χρησιμοποιήσαμε το ALARM δίκτυο που είναι σχετικά μεγάλο και αρκετά συνεκτικό. Για την αξιολόγηση της απόδοσης του συνολικού αλγορίθμου ξεκινάμε από ένα



Σχήμα 8.12: 'Ενα μέρος του γενετικού δικτύου του κύκλου του κυττάρου που πήραμε χρησιμοποιώντας γραμμική Gaussian κατανομή.

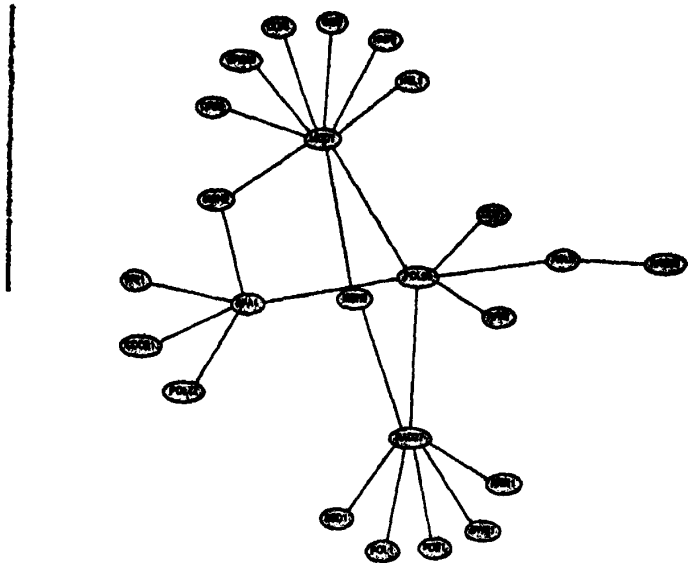


Σχήμα 8.13: Ένα μέρος του γενετικού δικτύου του κύκλου του κυττάρου που πήραμε χρησιμοποιώντας Radial Basis κατανομές.

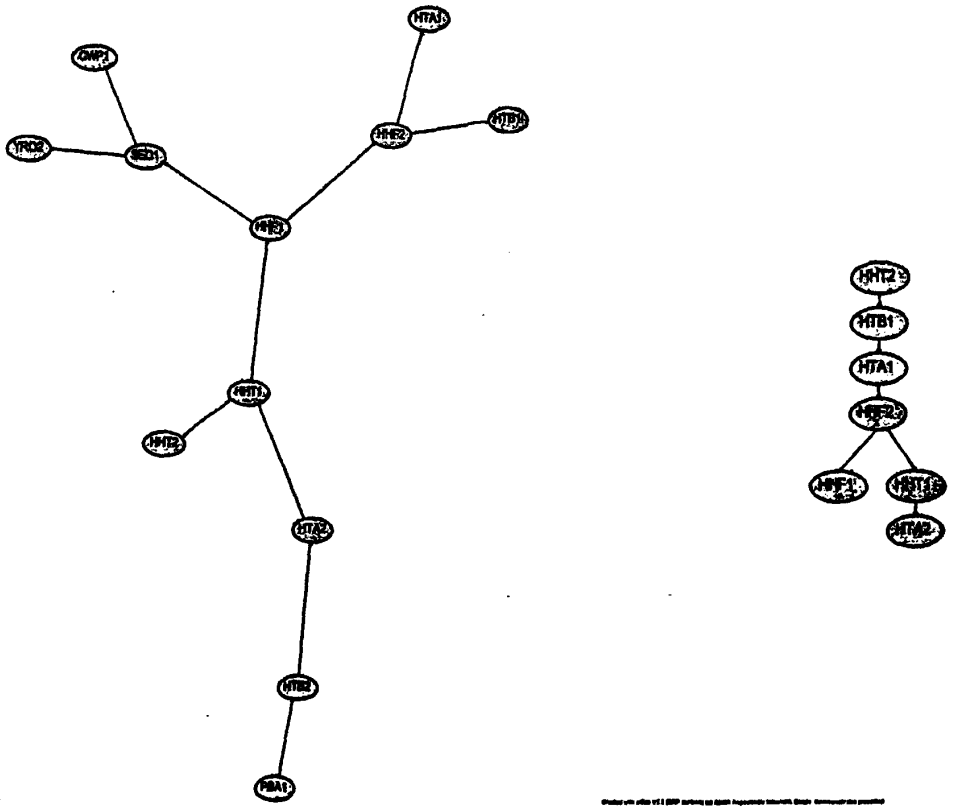


(α) (β)

Σχήμα 8.14: Η γειτονιά του γονιδίου CLN1 που είναι γνωστό ότι ενεργοποιείται στην αρχή του κύκλου του κυττάρου (α) Διακριτές μεταβλητές και (β)Συνεχείς μεταβλητές (Radial Basis).

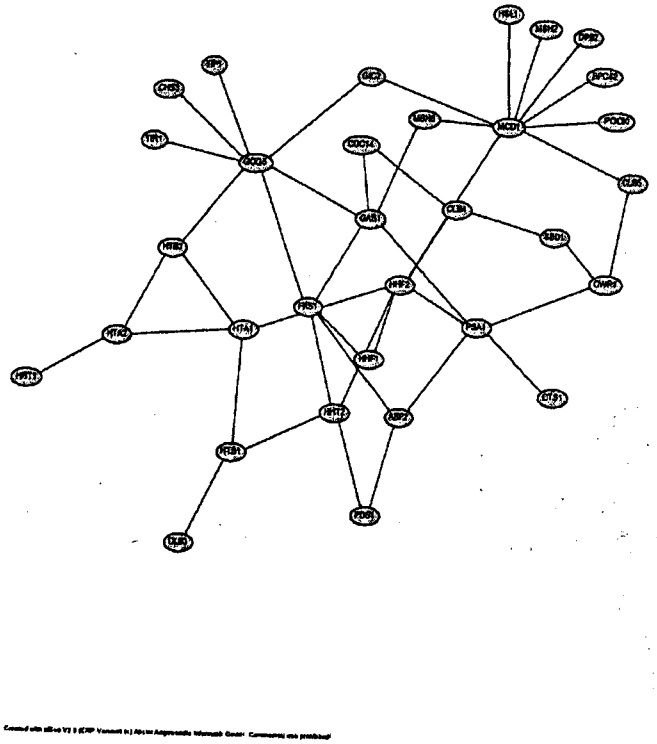


Σχήμα 8.15: Η γειτονιά του γονιδίου CLN1 για συνεχείς μεταβλητές (Γραμμική Gaussian κατανομή).

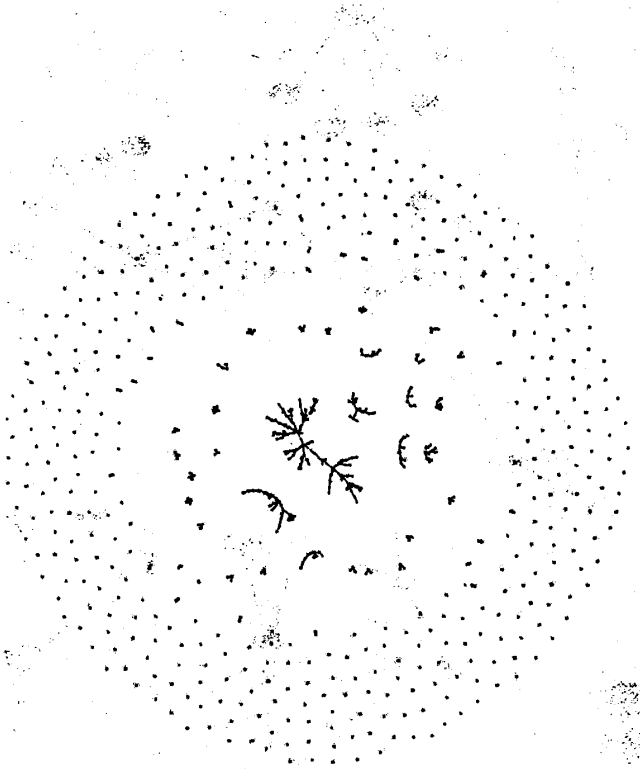


(α) (β)

Σχήμα 8.16: Η γειτονιά του γονιδίου HTA1 (α) Διακριτές μεταβλητές και (β)Συνεχείς μεταβλητές (Radial Basis).

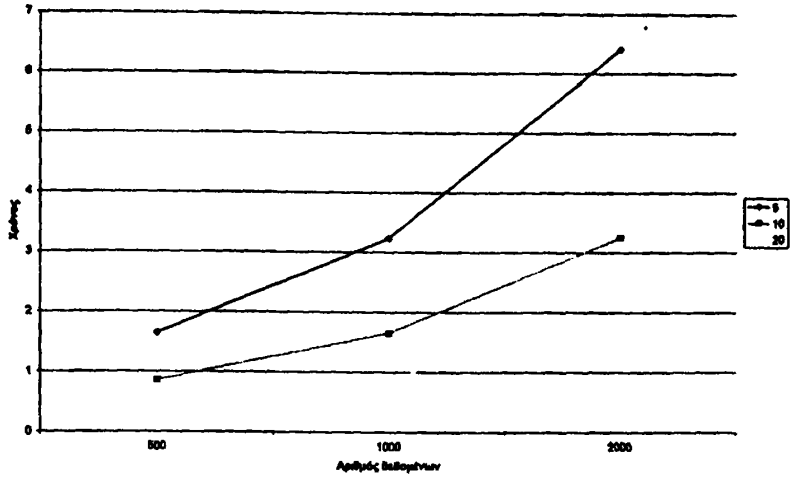


Σχήμα 8.17: Η γειτονιά του γονιδίου HTA1 για συνεχείς μεταβλητές (Γραμμική Gaussian κατανομή).



Σχήμα 8.18: Το δίκτυο για 800 γονίδια.

ΓΟΝΙΔΙΟ	ΠΑΤΕΡΕΣ	ΣΚΟΡ	ΓΟΝΙΔΙΟ	ΠΑΤΕΡΕΣ	ΣΚΟΡ
CLB5	CLB6 CLB2	-184.018799	CLN1	POL30 CLB2	-276.057404
HTB1	HTA1 HHF2	-224.821945	CIK1	CDC5 CDC21	-202.294968
TEC1	RME1 DBF2	-328.231110	SIC1	YRO2 CTS1	-220.363342
DBF4	DBF2	-222.290390	FAR1	YRO2 CLB2	-308.008362
HHT2	HHT1 HTA1	-201.520752	RFA1	POL30 CTS1	-283.895630
ASF2	DPB3	-206.732086	SPO12	DBF2	-187.640579
RAD17	RNR1	-129.854187	SPC42	MCD1	-164.882492
CHS3	CDC9	-230.800049	CWP1	PSA1 SED1	-379.234344
CLB4	HHF1	-171.648529	PCL1	SWE1	-147.812424
KIN3	CDC20	-306.692810	RAD27	CDC14	-256.909760
KRE6	FKS1	-245.097153	DBF2	CDC20	-251.577332
HTA2	HHT1 HTA1	-243.610870	GIC2	GOG5	-282.786438
RFA2	RNR1	-255.774734	POL2	POL30	-201.187744
SST2	STE2	-292.814453	RAD53	CLB6	-278.980042
CDC9	RAD27	-209.481552	CDC21	RAD27	-271.750793
MNN1	PRI2	-298.795441	ASF1	MSH6	-255.202209
HSL1	MSH6	-230.915283	MOB1	CDC20	-158.390396
SWE1	RAD27	-139.560806	POL1	MCD1	-238.995667
POL30	RAD27	-329.169098	PRI2	MCD1	-190.888138
RNR1	MCD1	-347.515289	CDC20	CDC5	-229.480896
MSH2	MSH6	-214.345444	GOG5	GAS1	-183.783112
HHT1	HHF2	-279.417175	MSH6	MCD1	-232.054291
HTA1	HHF2	-257.019470	CDC14		-144.234497
HHF2	HHF1	-197.817963	CLB1	CDC5	-229.087067
MCD1	POL30	-254.514771	CLB2	CDC5	-280.543274
CLB6	MSH6	-305.550903	SWI5	CLB1	-233.496719
PSA1	GAS1	-255.377533	POL12	POL30	-260.787842
ACE2	CLB1	-243.681183	STE2	MFA2	-240.585434
FKS1	GOG5	-255.018402	HHF1	PSA1	-337.690979
DPB2	MCD1	-228.404266	CDC47	DBF2	-319.509888
GAS1	MCD1	-300.393799	PMS1	RFA1	-192.224380
HPR5	FKS1	-201.522049	TIP1	GOG5	-381.517609
ASE1	CDC20	-177.098755	SWI4	CLB6	-188.702576
MFA2	FAR1	-268.860352	PDS1	MCD1	-251.448532
NUF1	CIK1	-132.690033	TIR1	POL1	-253.328262
RME1	YRO2 SIC1	-343.889343	RFA3	MSH2	-195.494690
CHS1	RNR1	-263.931488	CDC2	CLB6	-232.835999
KAR3	CLN1	-227.910370	CDC5	PSA1 RNR1	-343.583893
PRI1	RFA1	-115.610130	UNG1	CDC14	-154.437592
EGT2	POL30 FAR1	-379.215790	SED1	HHF1 YRO2	-306.080200
CTS1	HTA1 EGT2	-331.899170	KAR4	TEC1 AGA1	-206.993347
NUM1	RFA3	-205.622498	CLB3	CHS3	-120.567841
DPB3	RFA2	-170.222382	HTB2	HTA2 HHF2	-256.511169
AGA1	TIP1 SST2	-320.109772	YRO2	MNN1 HHF2	-414.080109

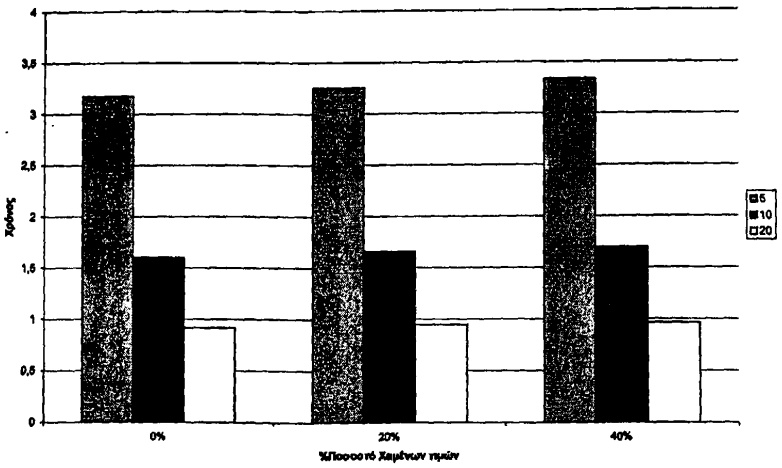


Σχήμα 8.19: Ο χρόνος σε σχέση με τον αριθμό των δεδομένων για μία επανάληψη στο ALARM δίκτυο (με 0.2% ελλιπείς τιμές) για 5,10,20 επεξεργαστές.

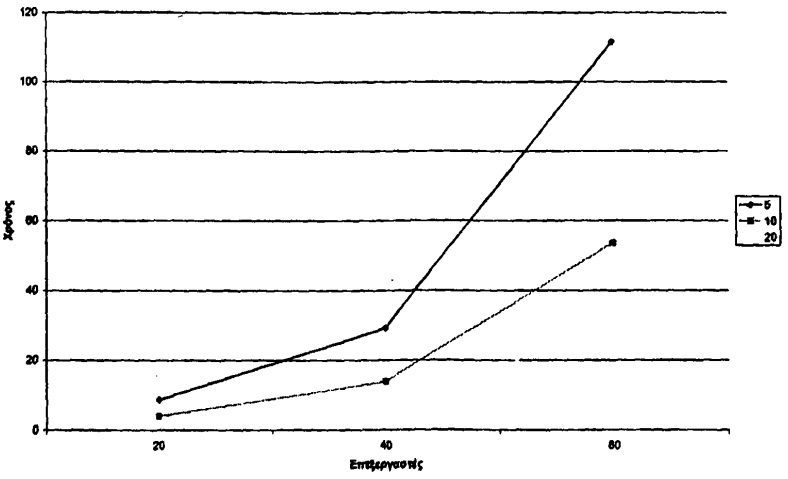
γράφημα με αριθμό ακμών περίπου ίσο με τον αριθμό των κόμβων. Σημειώνουμε ότι στο πρώτο μέρος όπου αξιολογούμε την απόδοση του συνολικού συστήματος, έχουμε διακριτές μεταβλητές.

Από τα Σχήματα 8.19, 8.20, 8.21, 8.22 μπορούμε να παρατηρήσουμε ότι επιτυγχάνεται καλή επιτάχυνση σε σχέση με όλους σχεδόν τους παράγοντες. Η παραλληλία στα επιμέρους τμήματα του συστήματος δηλαδή στην εύρεση των ML παραμέτρων και στην αξιολόγηση των υποψήφιων δικτύων εξαρτάται από διαφορετικούς παράγοντες. Στο πρώτο η απόδοση αυξάνεται, δηλαδή έχουμε μεγαλύτερη επιτάχυνση σε σχέση με τον αριθμό των επεξεργαστών όταν ο λόγος δεδομένων/επεξεργαστών μεγαλώνει. Στη δεύτερη περίπτωση αντίστοιχα πρέπει ο λόγος υποψηφίων δικτύων/επεξεργαστών να μεγαλώνει.

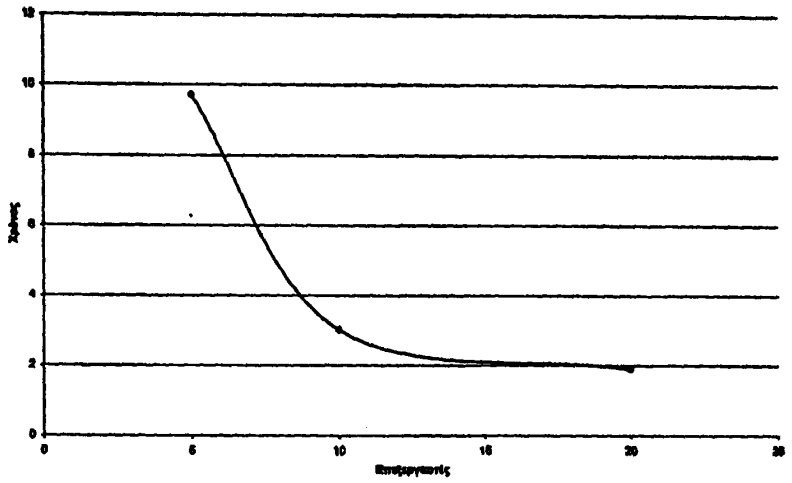
Στα Σχήματα 8.23(α), (β) και 8.24(α), (β) φαίνεται ο χρόνος εκτέλεσης για την εύρεση των παραμέτρων και την εύρεση του νέου δικτύου για συνεχή και διακριτά δίκτυα στην εξέλιξη του αλγορίθμου. Καθώς ο αλγόριθμος εύρεσης προχωράει το γράφημα γίνεται πιο συνεκτικό, νέες ακμές προστίθενται στο γράφημα. Είναι λογικό όσο πιο συνεκτικό γίνεται το γράφημα να αυξάνεται και το μέγεθος της μέγιστης κλίκας και όπως βλέπουμε και στα Σχήματα 8.23(α), (β) και όπως είναι γνωστό από την θεωρία ο αλγόριθμος συμπερασματολογίας είναι εκθετικός για δίκτυα με διακριτές μεταβλητές ως προς το πληθάριθμο της μεγαλύτερης κλίκας και έχει πολυωνυμική πολυπλοκότητα τρίτου βαθμού για δίκτυα με συνεχείς μεταβλητές. Στο Σχήμα 8.23 (για συνεχείς μεταβλητές) βλέπουμε σε κάποιο σημείο η αύξηση του χρόνου να είναι δραματική. Αυτό οφείλεται στο γεγονός ότι με συνεχείς γραμμικές κατανομές το γράφημα που πήραμε ήταν πολύ πιο συνεκτικό από αυτό με τις διακριτές κατανομές. Ενώ για διακριτές κατανομές ο μέγιστος αριθμός μεταβλητών μιας κλίκας ήταν μέχρι 4, για συνεχείς ξεπέρασε το 20, το οποίο είχε δραματική επίπτωση στην



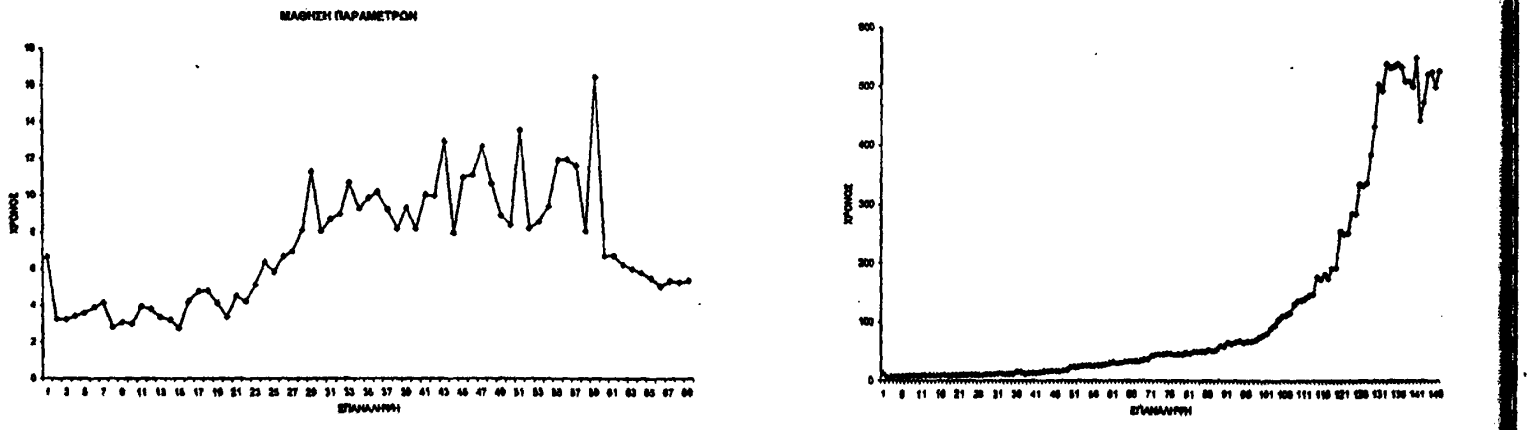
Σχήμα 8.20: Ο χρόνος σε σχέση με τον αριθμό των χαμένων τιμών για για μία επανάληψη στο ALARM δίκτυο (με 1000 δεδομένα) για 5,10,20 επεξεργαστές.



Σχήμα 8.21: Ο χρόνος σε σχέση με τον αριθμό των κόμβων του δικτύου για για μία επανάληψη σε τυχαίο δίκτυο (με 1000 δεδομένα και 0.2% ελλιπείς τιμές) για 5,10,20 επεξεργαστές.

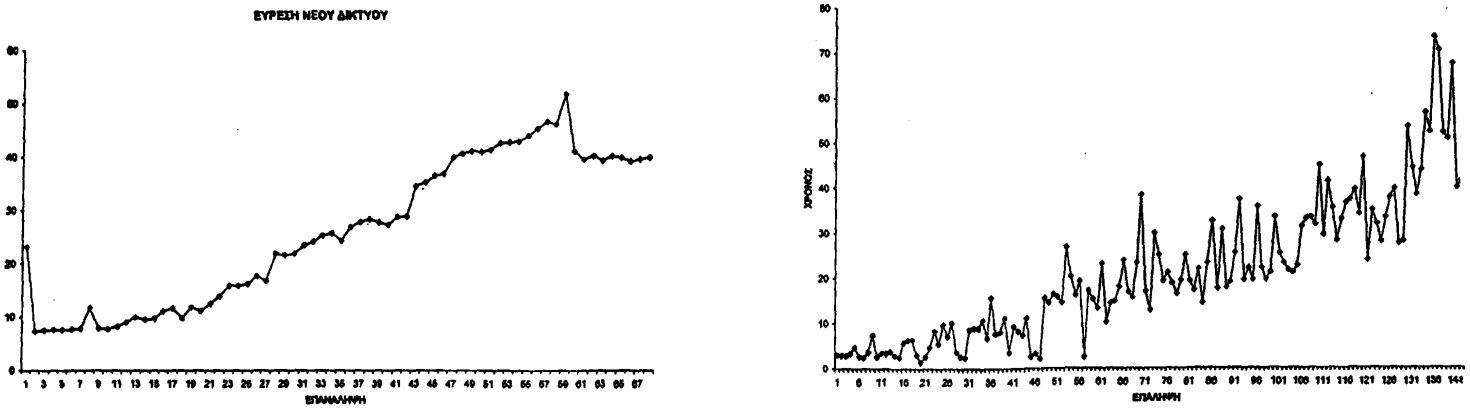


Σχήμα 8.22: Ο χρόνος σε μια μέση περίπτωση όπου έχουμε 20 κόμβους, 1000 δεδομένα και 20% ποσοστό χαμένων τιμών για 5,10,20 επεξεργαστές



(α) (β)

Σχήμα 8.23: Ο χρόνος εύρεσης των παραμέτρων για κάθε επανάληψη του αλγορίθμου (α) Για διακριτές μεταβλητές και (β) Για συνεχείς μεταβλητές.



(α) (β)

Σχήμα 8.24: Ο χρόνος εύρεσης νέου δικτύου για κάθε επανάληψη του αλγορίθμου (α) Για διακριτές μεταβλητές και (β) Για συνεχείς μεταβλητές.

επίδοση του αλγορίθμου. Το παράδοξο είναι ότι έχουμε θέσει ως μέγιστο όριο πατέρων το 5, οπότε μπορούμε να υποθέσουμε ότι είναι μειονέκτημα του αλγορίθμου κατασκευής του συνδετικού δέντρου.

8.7 Συμπεράσματα και μελλοντική εργασία

Σε αυτήν την διατριβή υλοποιήσαμε ένα σύστημα για την επεξεργασία δεδομένων από μικροσυστοιχίες DNA με στόχο την εύρεση εξαρτήσεων γονιδίων με την χρήση Bayesian δικτύων. Τα σημαντικότερα πλεονεκτήματα του συστήματος είναι η παραλληλία που μας επιτρέπει την επεξεργασία ενός μεγάλου αριθμού γονιδίων με Bayesian δίκτυα. Οι χρόνοι για ένα δίκτυο με 100 περίπου κόμβους ήταν αρκετά ενθαρρυντικοί, κυμαίνονται από 10 λεπτά ως κάποιες ώρες. Η μεγάλη διακύμανση οφείλεται στο γεγονός, όπως ήδη αναφέραμε, ότι σε πολλές περιπτώσεις τα γραφήματα γίνονται πολύ συνεκτικά με αποτέλεσμα την αύξηση της μέγιστης κλίκας και την ραγδαία επιβράδυνση του αλγορίθμου συμπερασματολογίας. Στα πλαίσια της μελλοντικής εργασίας μπορούμε να επιτύχουμε πιθανές βελτιώσεις τόσο στην παραλληλία όσο και στον ίδιο τον αλγόριθμο της συμπερασματολογίας.

- Η παραλληλοποίηση της κατασκευής του συνδετικού δέντρου η οποία αν και σε μικρά δίκτυα δεν είναι χρονοβόρα, σε μεγάλα δίκτυα απαιτεί αρκετό χρόνο.

- Θα μπορούσαμε από ένα σημείο και μετά που ο αλγόριθμος συνδετικού δέντρου δεν είναι πλέον αποδοτικός (με κάποιο κριτήριο) να καταφεύγουμε σε άλλους αλγορίθμους όπως στοχαστικούς (Monte Carlo).

- Όπως προαναφέραμε το υπολογιστικό κόστος του αλγορίθμου εξαρτάται κυρίως από τον αριθμό μεταβλητών στην μεγαλύτερη κλίκα. Σημαντική βελτίωση μπορούμε να έχουμε όταν ο αριθμός των μεταβλητών σε μία κλίκα είναι μικρός. Αυτό γίνεται στην κατασκευή του συνδετικού δέντρου και ειδικότερα στην τριγωνοποίηση του γραφήματος. Η βιβλιογραφία και η έρευνα πάνω σε αυτό το συγκεκριμένο κομμάτι το οποίο στην ουσία καθορίζει και τις κλίκες του δέντρου μας είναι ευρύτατη. Υπάρχουν πολλοί τρόποι και πολλές μεθοδολογίες για να πάρουμε όσο το δυνατόν λιγότερες προσθέσεις ακμών στο γράφημα. Ειδικά σε μεγάλα και αρκετά συνεκτικά γραφήματα το κέρδος μπορεί να είναι σημαντικό.

- Μία πιθανή βελτίωση θα ήταν στην περίπτωση που το γράφημα μας είναι μη συνεκτικό και αποτελούταν από μεγάλο αριθμό υπογραφημάτων οπού μπορούσαμε να αναθέσουμε σε κάθε επεξεργαστή την εύρεση των παραμέτρων καθε τέτοιου υπογραφήματος καθώς οι παράμετροι του θα είναι ανεξάρτητοι από του υπόλοιπου υπογραφήματος. Αλλά και πάλι από ένα σημείο και μετά που το γράφημα γίνεται πιο συνεκτικό καθώς προστίθενται ακμές η απόδοση θα πέφτει. Μπορούμε επίσης να χρησιμοποιήσουμε και τις δύο παραπάνω στρατηγικές ή μία δυναμική κατανομή επεξεργαστών σε υπογραφήματα που ο αριθμός μεταβλητών της μεγαλύτερης κλίκας να είναι σχετικά μεγάλος.

Στα πειράματα διαπιστώσαμε την υπεροχή του SEM σε σύγκριση με την ανάκτηση χαμένων τιμών από κλασικές μεθόδους όπως ο K-NN. Άναφέραμε πως δεν κάναμε πειράματα σε υβριδικά δίκτυα επειδή τα δεδομένα δεν ήταν επαρκή, γεγονός που προκαλεί αριθμητικά λάθη. Πρέπει ως μελλοντική εργασία να βρούμε πληρέστερη βάση δεδομένων ώστε και τα

υβριδικά δίκτυα να δοκιμάσουμε αλλά και πιο τεκμηριωμένα αποτελέσματα να πάρουμε για τα γενετικά δίκτυα. Μία ακόμα επέκταση του συστήματος όπως έχουμε αναφέρει θα ήταν στα υβριδικά δίκτυα να μπορούν οι διακριτές μεταβλητές να έχουν συνεχείς πατέρες. Τέλος θα πρέπει να δούμε και την απόδοση του αλγορίθμου SEM χρησιμοποιώντας κάποια άλλη προσέγγιση για το αναμενόμενο σκορ εκτός από την γραμμική (5.17).

ΑΝΑΦΟΡΕΣ

[1] S. Draghici. *Data Analysis Tools for DNA Microarrays*. Chapman Hall /CRC Press, 2003.

[2] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. In *RECOMB '00: Proceedings of the fourth annual international conference on Computational molecular biology*, pages 127–135, New York, NY, USA, 2000. ACM Press.

[3] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95(25):14863–14868, 1998.

[4] M. Kloster, C. Tang, and N.S. Wingreen. Finding regulatory modules through large-scale gene-expression data analysis. *Bioinformatics*, 21(7):1172–1179, 2005.

[5] S. Weng, Q. Dong, R. Balakrishnan, K. Christie, M. Costanzo, K. Dolinski, S. S. Dwight, S. Engel, D. G. Fisk, E. Hong, L. Issel-Tarver, A. Sethuraman, C. Theesfeld, R. Andrada, G. Binkley, C. Lane, M. Schroeder, D. Botstein, and J. M. Cherry. Saccharomyces genome database (sgd) provides biochemical and structural information for budding yeast proteins. *Nucl. Acids Res.*, 31(1):216–218, 2003.

[6] N. Friedman. Learning belief networks in the presence of missing values and hidden variables. In *Proc. 14th International Conference on Machine Learning*, pages 125–133. Morgan Kaufmann, 1997.

[7] J. H. Kim, I. S. Kohane, and L. Ohno-Machado. Visualization and evaluation of clusters for exploratory analysis of gene expression data. *J. of Biomedical Informatics*, 35(1):25–36, 2002.

[8] G. Wong C. Eero P. Toronen, M. Kolehmainen. Analysis of gene expression data using self-organizing maps. *FEBS Letters*, 451:v, 5 1999.

[9] E. Hartuv, A. Schmitt, J. Lange, S. Meier-Ewert, H. Lehrachs, and R. Shamir. An algorithm for clustering cDNAs for gene expression analysis. In *RECOMB*, pages 188–197, 1999.

[10] D. C. Weaver, C. T. Workman, and G. D. Stromo. Modeling regulatory networks with weight matrices. *PSB*, 1999.

[11] W. Spears. Simulated annealing for hard satisfiability problems, 1993.

[12] P. D'haeseleer, X. Wen, S. Fuhrman, and R. Somogyi. Linear modeling of mrna expression levels during cns development and injury, 1999.

[13] T. Chen, H. He, and G. Church. Modeling gene expression with differential equations, 1999.

[14] M. Learning and Dietterich. Machine learning, 1990.

[15] S. A. Kauffman. *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, May 1993.

[16] H. Lhdesmki, I. Shmulevich, and O. Yli-Harja. On learning gene regulatory networks under the boolean network model. *Mach. Learn.*, 52(1-2):147–167, 2003.

[17] S. Liang, S. Fuhrman, and R. Somogyi. Reveal: a general reverse engineering algorithm for inference of genetic network architectures, 1998.

[18] T. Akutsu. Identification of genetic networks from a small number of gene expression patterns under the boolean network model, 1999.

[19] A. Silvescu and V. Honavar. Temporal boolean network models of genetic networks and their inference from gene expression time series. *Complex Systems*, 13:54–70, 2001.

[20] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Mol. Biol. Cell*, 9(12):3273–3297, 1998.

[21] P. Spirtes, C. Glymour, and R. Scheines. Constructing bayesian network models of gene expression networks from microarray data, 2000.

[22] T. Goto S. Aburatani K. Tashiro S. Kuhara S. Miyano S. Imoto, K. Sunyong. Bayesian Network and Nonparametric Heteroscedastic Regression for Nonlinear Modeling of Genetic Network. *IEEE Computer Society Bioinformatics Conference (CSB'02)*, page 219, 2002.

[23] K. Murphy and S. Mian. Modelling gene expression data using dynamic bayesian networks, 1999.

[24] T. Jaakkola A. Hartemink, D. Gifford and R. Young. Using Graphical Models and Genomic Expression Data to Statistically Validate Models of Genetic Regulatory Networks. In *Sixth Pacific Symposium on Biocomputing*, 2001.

[25] I. M. Ong, J. D. Glasner, and D. Page. Modelling regulatory pathways in E. coli from time series expression profiles. *Bioinformatics*, 18(90001):241S–248, 2002.

[26] B. E. Perrin, L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet, and F. d'Alche Buc. Gene networks inference using dynamic Bayesian networks. *Bioinformatics*, 19(90002):138ii–148, 2003.

[27] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[28] S. Oba, M. Sato, I. Takemasa, M. Monden, K. Matsubara, and S. Ishii. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16):2088–2096, 2003.

[29] H. Kim, G. H. Golub, and H. Park. Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics*, 21(2):187–198, 2005.

[30] G. Michaels, D. Carr, M. Askenazi, S. Fuhrman, X. Wen, and R. Somogyi. Cluster analysis and data visualization of large-scale gene expression data, 1998.

[31] R. Herwig, A. J. Poustka, C. Muller, C. Bull, H. Lehrach, and J. O'Brien. Large-Scale Clustering of cDNA-Fingerprinting Data. *Genome Res.*, 9(11):1093–1105, 1999.

[32] F. De Smet, J. Mathys, K. Marchal, G. Thijs, B. De Moor, and Y. Moreau. Adaptive quality-based clustering of gene expression profiles. *Bioinformatics*, 18(5):735–746, 2002.

[33] J. A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, Inc., New York, NY, USA, 1975.

[34] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[35] M. H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill Book Company, wcl edition edition, 2004.

[36] G. F. Cooper and E. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Mach. Learn.*, 9(4):309–347, 1992.

[37] H. Akaike. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[38] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[39] D. Heckerman, A. Mamdani, and M.P. Wellman. Real-world applications of bayesian networks. *Commun. ACM*, 38(3):24–26, 1995.

[40] W. L. Buntine. Theory refinement of Bayesian networks. In *Uncertainty in Artificial Intelligence*, 1991.

[41] D. Heckerman and D. Geiger. Learning Bayesian Networks. Technical Report MSR-TR-95-02, Microsoft Research, Redmond, WA, December 1994.

[42] D. Chickering. Learning bayesian networks is np-complete, 1995.

[43] D. S. Johnson, C. H. Papadimtriou, and M. Yannakakis. How easy is local search? *J. Comput. Syst. Sci.*, 37(1):79–100, 1988.

[44] D. Geiger, D. Heckerman, and C. Meek. Asymptotic model selection for directed networks with hidden variables. pages 283–290.

[45] D. B. Rubin G. W. Imbens. Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics*, 25(1):305–327, , note = , abstract = , keywords = , source = , 1997.

[46] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. pages 452–472, 1990.

[47] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–779, ???? 1995.

[48] N. Friedman. The bayesian structural em. *Uncertainty in Artificial Intelligence (UAI)*, 14:29–31, July 1998.

[49] J. Binder, D. Koller, S. J. Russell, and K. Kanazawa. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29(2-3):213–244, 1997.

[50] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. pages 415–448, 1990.

[51] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm.

[52] S. L. Lauritzen. The em algorithm for graphical association models with missing data. *Comput. Stat. Data Anal.*, 19(2):191–201, 1995.

[53] P. Heggernes. Minimal triangulations of graphs: A survey. Technical report, Department of Informatics, University of Bergen.

[54] A. Berry, J. R. S. Blair, and P. Heggernes. Maximum cardinality search for computing minimal triangulations. In *WG '02: Revised Papers from the 28th International Workshop on Graph-Theoretic Concepts in Computer Science*, pages 1–12, London, UK, 2002. Springer-Verlag.

[55] F. V. Jensen F. Jensen. Optimal junction trees. *Uncertainty in Artificial Intelligence*, pages 29–31, July 1994.

[56] F.V. Jensen. Junction Trees and Decomposable Hypergraphs. Research report, JUDEX data-systemer A/S, 1988.

[57] M. Gondran, M. Minoux, and S. Vajda. *Graphs and algorithms*. John Wiley & Sons, Inc., New York, NY, USA, 1984.

[58] P. P. Shenoy and G. Shafer. Axioms for probability and belief-function propagation. *Uncertainty in artificial intelligence*, IV:169–98, 1990.

[59] S. L. Lauritzen and F. V. Jensen. Local computation with valuations from a commutative semigroup. *Annals of Mathematics and Artificial Intelligence*, 21(1):51–69, 1997.

[60] C. Huang A. Darwiche. Inference in belief networks: A procedural guide. *International Journal of Approximate Reasoning*, 11, 1994.

[61] S. L. Lauritzen. Propagation of probabilities, means, and variances in mixed graphical association models. *Journal of the American Statistical Association*, 87(420):1098–1108, 1992.

[62] S. L. Lauritzen. Graphical models. Technical report, OUP, 1996.

[63] S. Lauritzen and F. Jensen. Stable local computation with conditional gaussian distributions.

[64] S.L. Lauritzen and N. Wermuth. Mixed interaction models. Technical report, Institute for Electronic Systems, Aalborg University, 1984.

[65] G. Shafer and J. Pearl, editors. *Readings in uncertain reasoning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990.

[66] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Inc., New York, NY, USA, 1995.

[67] S. Konishi, T. Ando, and S. Imoto. Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika*, 91(1):27–43, 2004.

[68] K. P. Murphy. A variational approximation for bayesian networks with discrete and continuous latent variables. pages 457–466.

Βιογραφικο

Ο Ρήγας Γεώργιος γεννήθηκε το 1981. Αποφοίτησε από το γενικό λύκειο Αριδαίας το 1999 και την ίδια χρονιά εισάχθηκε στο Πανεπιστήμιο Ιωαννίνων στο τμήμα Πληροφορικής και αποφοίτησε το 2003. Στην πτυχιακή του εργασία ασχολήθηκε με "επίλυση διαφορικών εξισώσεων με νευρωνικά δίκτυα σε παράλληλους υπολογιστές".