“Ελεγχος Παραμέτρων Ποιότητας με Μπαεζιανά Δίκτυα .Μια εφαρμογή σε μεγάλο Ευρωπαϊκό Οργανισμό Κατασκευής Τηλεπικοινωνιακού Υλικού”

από τον
Μαθιουδάκη Αδάμ

Μια πτυχιακή εργασία που εκπονήθηκε στα πλαίσια της
μερικής εκπλήρωσης των απαιτήσεων για την απόκτηση του πτυχίου:
“Μηχανικού Παραγωγής και Διοίκησης”

Πολυτεχνείο Κρήτης

2003

Εγκρίθηκε από _____

Επιβλέπων καθηγητής

Λουκάς Τσιρώνης _____

Βασίλης Μουστάκης _____

Ευάγγελος Γρηγορούδης _____

Ημερομηνία: 13 Οκτωβρίου 2003

Ευχαριστίες

Σ’ αυτό το σημείο, θα ήθελα να ευχαριστήσω τους γονείς μου,οι οποίοι με βοήθησαν σε όλη μου αυτή την προσπάθεια και διαδρομή. Όπως επίσης, τους καθηγητές μου κ. Λουκά Τσιρώνη και Βασίλη Μουστάκη, οι οποίοι με βοήθησαν στην πορεία αυτής μου της προσπάθειας.

“ Change is eternal nothing ever change”

ΙΜΜΑΝΟΥΕΛ ΒΑΛΛΕΡΣΤΕΪΝ

Περιεχόμενα

Κεφάλαιο 1ο

Κεφάλαιο 2ο

Κεφάλαιο 3ο

Κεφάλαιο 4ο

Κεφάλαιο 5ο

Παραρτήματα

Περίληψη

Σ ’αυτή την ερευνητική εργασία μελετήθηκε η διαδικασία παραγωγής ενός isdn modem και η οποία παρουσιάζει μεγάλο ποσοστό μη λειτουργικών προιόντων.Σκοπός μας είναι σε πρώτο στάδιο να γίνει στατιστική ανάλυση των δεδομένων τα οποία έχουν ήδη καταγραφεί.Στη συνέχεια και με τη βοήθεια διαφόρων λογισμικών τα οποία κάνουν χρήση της θεωρίας του ***BAYES*** και σε συνδυασμό με τη βάση δεδομένων θα δημιουργήθουν τα λεγόμενα ***Bayesian networks(BNs)*** ή ***Belief networks*** μέσω των οποίων θα γίνει η εξαγωγή χρήσιμων συμπερασμάτων με τελικό σκοπό την ελαχιστοποίηση των σφαλμάτων στην παραγωγική αλυσίδα και επομένως την μείωση των προιόντων τα οποία δεν πληρούν τα χαρακτηριστικά ποιότητας.Η δομή της εργασίας εν συντομία είναι η ακόλουθη:

Στο **κεφάλαιο 1** γίνεται αναφορά στο χώρο που εφαρμόσθηκε η προτεινόμενη μεθοδολογία της πτυχιακής. Αφορά τη διαδικασία παραγωγής συσκευών modem. Παρουσιάζεται το διάγραμμα ροής της διαδικασίας και προσδιορίζονται τα χαρακτηριστικά ποιότητας στα οποία στηρίζεται ο έλεγχος ποιότητας της διαδικασίας και του προϊόντος καθώς και πίνακας με τα ελαττώματα που παρουσιάστηκαν συνοδευόμενος με μια σύντομη περιγραφή τους. .Στο **κεφάλαιο 2** επιχειρήται μια σύντομη εισαγωγή στην θεωρίας των μπαυζιανών δικτύων .Στο **κεφάλαιο 3** υπάρχει μια σύντομη παρουσίαση των δεδομένων και των δυνατοτήτων των εμπορικών προγραμμάτων των οποίων θα γίνει χρήση.Στο **κεφάλαιο 4** αποτυπώνεται η όλη διαδικασία που γίνεται για την εξαγωγή αποτελεσμάτων,ενώ στο τελευταίο κεφάλαιο παρουσιάζονται τα αποτελέσματα και ακολουθεί και η κριτική των αποτελεσμάτων έτσι ώστε να καταλήξουμε στις απαραίτητες ενέργειες βελτίωσης της γραμμής παραγωγής.Επίσης το τελευταίο κομμάτι της εργασίας αποτελείται από τα παραρτήματα ***A*** και ***B***

Κεφάλαιο 1^0

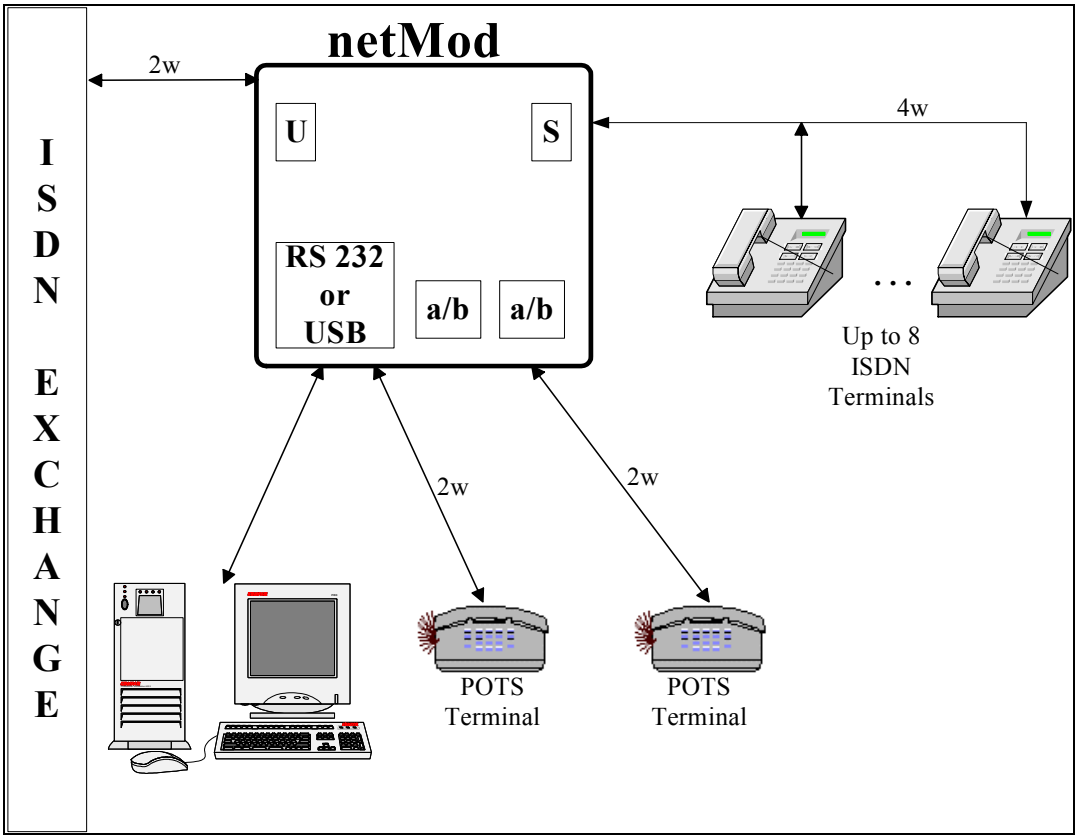
1.0.1 Εισαγωγή

Σκοπός αυτου του κεφαλαίου είναι η περιγραφή της γραμμής παραγωγής ενός οργανισμού κατασκευής τηλεπικοινωνιακού εξοπλισμού και πιο συγκεκριμένα ενός ISDN modem με την ονομασία NETMOD καθώς και των χαρακτηριστικών του εν λόγου προιόντος. Εκτός από την παρουσίαση των σταδίων παραγωγής και των σταθμών ελέγχου σκοπός του κεφαλαίου είναι να προσδιορίσει τα χαρακτηριστικά ποιότητας τα οποία είναι υπεύθυνα για το επίπεδο ποιότητας της διαδικασίας και του προϊόντος.Στη συνέχεια δίνεται μια σύντομη περιγραφή των δυνατοτήτων του ***isdn modem*** ,το οποίο μελετάμε.

NetMod

Το NetMod παρέχει τη δυνατότητα σύνδεσης τερματικών συσκευών τύπου ISDN και απλών αναλογικών συσκευών και τη δυνατότητα σύνδεσης προσωπικού υπολογιστή σε δίκτυα όπως το Internet χωρίς τη χρήση επιπλέον εξοπλισμού. Η σύνδεση μέχρι οκτώ ISDN τερματικών συσκευών γίνεται μέσω του S-bus, των απλών αναλογικών συσκευών, ενώ των απλών αναλογικών συσκευών γίνεται μέσω δύο αναλογικών θυρών, ενώ η σύνδεση με τον προσωπικό υπολογιστή γίνεται μέσω μιας ασύγχρονης σειριακής θύρας (RS232).

Το netMod προσφέρει τη δυνατότητα σύνδεσης τερματικών συσκευών ISDN, συσκευών τύπου POTS (αναλογικά τηλέφωνα, Fax) και μέσω PC πρόσβαση σε δίκτυα χωρίς επιπλέον εξοπλισμό.

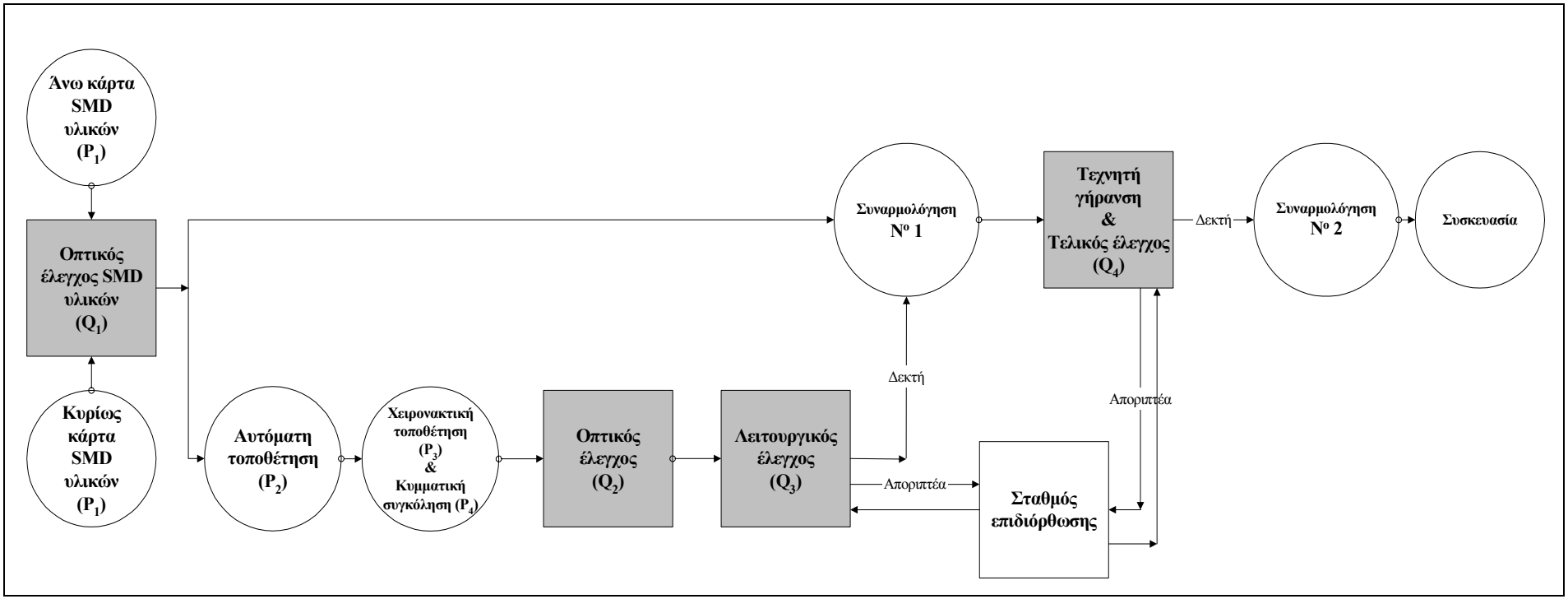


Σχήμα 1.1: Η σύνδεση του NT με τις τερματικές συσκευές

1.1 Η διαδικασία παραγωγής του netMod

Η γραμμή παραγωγής του NT αποτελείται από 12 στάδια εκ των οποίων τα τέσσερα είναι οι σταθμοί ελέγχου του, όμως επειδή τελικά συγχωνεύονται ορισμένοι σταθμοί η γραμμή παραγωγής αποτελείται από 4 σταθμούς παραγωγής (P1-P4) και 4 σταθμούς ελέγχου (Q1-Q4). Μεταξύ του λειτουργικού και του τελικού ελέγχου υπάρχει ένα επιπλέον στάδιο που είναι ο σταθμός επιδιόρθωσης. Στο σταθμό αυτό επιδιορθώνονται τα ελαττωματικά προϊόντα που προκύπτουν από τον λειτουργικό και τον τελικό έλεγχο καθώς και αυτά που δεν έχουν εντοπισθεί από τους προηγούμενους σταθμούς ελέγχου. Η διαδικασία παραγωγής παρουσιάζεται στο σχήμα1.2. Πρόκειται για μια γραμμή συναρμολόγησης στην οποία το μόνο μη-αυτοματοποιημένο στάδιο είναι η χειρονακτική τοποθέτηση μεγάλων συμβατικών υλικών.

Με κύκλο παρουσιάζονται τα στάδια παραγωγής και συναρμολόγησης των υλικών και με τετράγωνο οι σταθμοί ελέγχου.



Σχήμα1.2: Διάγραμμα ροής διαδικασίας παραγωγής modem

1.1.1 Φάσεις παραγωγής αναλυτικά

Το προϊόντα αποτελείται από δύο πλακέτες την κυρίως πλακέτα και την άνω πλακέτα που τοποθετείται πάνω στην κυρίως

1.1.1.1 Τοποθέτηση SMD υλικών στην άνω κάρτα (P_1)

Τοποθετούνται υλικά και στις δύο πλευρές της κυρίως πλακέτας, ταυτόχρονα, από τις μηχανές αυτόματης τοποθέτησης. Πάνω στα «πατάκια» της πλακέτας προστίθεται μια σταθεροποιητική ουσία που ονομάζεται «πάστα». Η πάστα σταθεροποιεί τα υλικά πάνω στην πλακέτα. Τα συγκεκριμένα υλικά λέγονται επιφανειακά (*Surface Mount Device-SMD*) λόγω ότι οι ακροδέκτες τους δεν διαπερνούν την πλακέτα. Ακολούθως η πλακέτα οδηγείται στο επόμενο στάδιο το οποίο είναι το στάδιο συγκόλλησης των υλικών με εφαρμογή θερμοκρασίας.

1.1.1.2 Τοποθέτηση SMD υλικών στην κυρίως κάρτα (P_1)

Για την πάνω πλευρά της κυρίως πλακέτας οι φάσεις παραγωγής είναι παρόμοιες με το προηγούμενο στάδιο. Στην κυρίως πλακέτα όμως υπάρχουν και άλλες κατηγορίες υλικών. Τα υλικά αυτά έχουν πιο μεγάλο μέγεθος από τα επιφανειακά και οι ακροδέκτες τους διαπερνούν την πλακέτα. Ονομάζονται συμβατικά υλικά (*radial & axial*).

1.1.1.3 Οπτικός έλεγχος SMD υλικών (Q_1)

Είναι ο πρώτος σταθμός ελέγχου της διαδικασίας.,όπου λαμβάνει χώρα οπτικός έλεγχος από τους αρμόδιους εργαζομένους και στις δύο πλευρές και στις δύο πλακέτες. Αφορά έλεγχο ποιότητας του προϊόντος στα στάδια στα οποία έχει ήδη περάσει. Τα χαρακτηριστικά ποιότητας τα οποία ελέγχονται στο στάδιο αυτό είναι:

1. Σωστή τοποθέτηση πάστας στα πατάκια (*pats*), αν το πατάκι γυαλίζει σημαίνει ότι έγινε σωστή τοποθέτηση αν όχι τότε η κόλλα δεν τοποθετήθηκε σωστά. Ύφος πάστας 170-220 μm. Πρέπει να υπάρχει σύμπτωση πατάκι-κόλλας

Σωστή τοποθέτηση υλικών

2. Απουσία υλικού, εξετάζεται η περίπτωση που κάποιο υλικό δεν τοποθετήθηκε
3. Σωστή συγκόλληση, οι βέλτιστες συνθήκες συγκόλλησης γίνονται σε θερμοκρασία: 210-225^0 C και χρόνο 45΄΄-90 ΄΄.

1.1.1.4 Αυτόματη τοποθέτηση συμβατικών υλικών (P_2)

Στο στάδιο αυτό τοποθετούνται αυτόματα τα συμβατικά υλικά. Αρχικά γίνεται αυτόματη τοποθέτηση των πλακετών στο χώρο κατεργασίας και κατόπιν πραγματοποιείται η τοποθέτηση κάποιου υλικού του επονομαζόμενου και «μάσκα» και το οποίο προστατεύει τα υλικά κατά το στάδιο της κυματικής συγκόλλησης ονομάζεται..Τέλευταία επεξεργασία αυτού του σταθμού παραγωγής είναι η προδιαμόρφωση των υλικών, δηλαδή οι μεγάλοι σε μήκος ακροδέκτες κόβονται και λυγίζονται για την επίτευξη καλύτερης συγκόλλησης.

1.1.1.5 Χειρονακτική τοποθέτηση (P_3)

Οι εργαζόμενοι τοποθετούν, χειρονακτικά, τα «μεγάλα» μεγέθους αντικείμενα (πυκνωτές, πηνία) τα οποία λόγω των διαστάσεών τους δεν μπορούν να τοποθετηθούν από την μηχανή. Στην συνέχεια η πλακέτα οδηγείται στην κυματική συγκόλληση όπου υλικά συγκολλούνται.

1.1.1.6 Κυματική συγκόλληση (Wave soldering) (P_4)

Στη φάση αυτή η κάρτα είναι πλήρης από τα υλικά της. Ακολουθεί ο καθαρισμός κάρτας (fluxing) από τυχών ακαθαρσίες ,η προθέρμανση της πλακέτας από τους τρεις προθερμαντήρες με εφαρμογή θερμοκρασίας (προθέρμανση, τρεις προθερμαντήρες (1ος. Κυκλοφορία θερμού αέρα 150^0 C, 2ος. Infrared 340^0 C, 3ος. Infrared 360^0 C), ταχύτητα ταινίας: 1,8m/min, δοχείο κόλλας: 340^0 C, $\pm 5^0$ C, χρόνος επαφής υλικού-κόλλας: 3-4 sec) και τέλος η πλακέτα οδηγείται στην κυματική συγκόλληση(μπάνιο)

Η τοποθέτηση των υλικών της πάνω πλευράς, στην κυρίως πλακέτα είναι αυτόματη και διενεργείται ταυτόχρονα.Έπειτα ακολουθεί η συγκόλληση των υλικών. Η πλακέτα περνάει εφαπτομενικά από λεκάνη με κόλληση θερμοκρασίας $360\ ^0$ C, η πλακέτα μόλις που εφάπτεται της επιφάνειας της κόλλησης και η κόλληση συγκρατείται στα πατάκια της πλακέτας. Ακολουθεί η διαδικασία reflow για την επίτευξη της συγκόλλησης των υλικών.

Οπτικός Έλεγχος (Q_2)

Είναι ο δεύτερος σταθμός ελέγχου ποιότητας. Διενεργείται οπτικός έλεγχος στην άνω πλευρά των συμβατικών υλικών και στην κάτω πλευρά των επιφανειακών υλικών. Τα χαρακτηριστικά ποιότητας που ελέγχονται είναι:

Κάτω πλευρά:

- Παρουσία όλων των υλικών
- Εξετάζεται η σωστή τοποθέτηση των υλικών, αν είναι σωστή η φορά των υλικών.
- Επίτευξη σωστή συγκόλλησης

Πάνω πλευρά:

- Έλεγχος υλικών χειρονακτικής και αυτόματης
- Σωστή φορά των υλικών
- Παρουσία υλικών
- Ξεμασκάρισμα
- Επικόλληση bar code

1.1.1.7 Λειτουργικός έλεγχος ποιότητας (Q_3)

Τα ποιοτικά χαρακτηριστικά που παρακολουθούνται σ' αυτό το στάδιο είναι:
NPC Έλεγχος τροφοδοσίας δηλαδή αν η τάση της κάρτας είναι μέσα στα επιτρεπτά όρια που έχει ορίσει η εταιρεία,και η οποία πικοίλει ανάλογα με τον τόπο εξαγωγής. Συγκεκριμένα οι μετρήσεις περιλαμβάνουν:

- Κατά τη διαδικασία αυτή πραγματοποιούνται τα test ελέγχου του τροφοδοτικού και του U-interface που περιλαμβάνουν τα εξής:
  i. Μετρήσεις των τάσεων εξόδου χωρίς φορτίο
  ii. Μετρήσεις των τάσεων εξόδου με πλήρες φορτίο
  iii. Έλεγχος του current limit
  iv. Έλεγχος του U-interface και της σηματοδοσίας προς αυτό
  v. Έλεγχος του S-interface και της σηματοδοσίας προς αυτό

- Αν το προϊόν είναι αλάνθαστο τότε η κυρίως κάρτα ενώνεται με την άνω pass ενώνεται με την upper στο επόμενο στάδιο συναρμολόγησης
- Αν ανιχνευθεί κάποιο ελάττωμα τότε οδηγείται εκ νέου στον σταθμό επιδιόρθωσης, και περνάει ξανά λειτουργικό έλεγχο.

Η δυναμικότητα του λειτουργικού ελέγχου είναι 50 κάρτες την ώρα.

1.1.1.8 Σταθμοί συναρμολόγησης

Αφού προηγουμένως έχει περάσει από τον λειτουργικό έλεγχο το προιόν οδηγείται στο σταθμό συναρμολόγησης. Πρώτα συναρμολογούνται οι δύο πλακέτες, η κυρίως με την άνω, έπειτα τοποθετούνται τα καλώδια τροφοδοσίας και το «βραχύκλωμα». Βραχυκύκλωμα ονομάζεται η δημιουργία διόδου σε ορισμένο σημείο της πλακέτας για να ολοκληρωθεί ένα κύκλωμα.και το οποίο δεν πρέπει να υπάρχει κατά το λειτουργικό έλεγχο, επομένως τοποθετείται μετά απ' αυτόν πρέπει όμως να υπάρχει στο τελικό προιόν.

1.1.1.9 Τεχνητή γήρανση – Τελικός έλεγχος (Q_4)

Η τεχνητή γήρανση είναι τεχνική καταπόνησης για να ελεγχθεί ο χρόνος ζωής και η αντοχή του προϊόντος στο χρόνο. Αυτός ο έλεγχος πραγματοποιείται με την εισαγωγή και θέρμανση του προιόντος σε φούρνο στους 50^0 C
Χαρακτηριστικά ποιότητας

- Έλεγχος σηματοδοσίας. Ελέγχεται η δίοδος σωστών σημάτων προς τα κανάλια του κέντρου. Κέντρο είναι μια κονσόλα η οποία χρησιμοποιείται ως προσομοιωτής του προϊόντος.
- Μετρήσεις κυκλωμάτων φωνής. Μέτρηση στάθμης θορύβου όλων των κυκλωμάτων σε (db).

- Μετρήσεις U activation. Έλεγχος ενεργοποίησης ή όχι της γραμμής δικτύου (U γραμμή)
- Μετρήσεις S activation. Έλεγχος ενεργοποίησης ή όχι των γραμμών εξόδου

(S γραμμή)

- Μετρήσεις σειριακής. Έλεγχος ενεργοποίησης ή όχι της σειριακής γραμμής.

Αν η κάρτα είναι δεκτή τότε οδηγείται στην συναρμολόγηση 2 κατά την οποία το προϊόντος τοποθετείται στο πλαστικό κουτί του και παίρνει την τελική του μορφή. Αν ανιχνευθεί κάποιο ελάττωμα τότε η πλακέτα οδηγείται στον σταθμό επιδιόρθωσης. Η διάρκεια του τελικού ελέγχου και της γήρανσης είναι 10 ώρες και η χωρητικότητα του ελέγχου είναι 585 κάρτες.

Στο παράρτημα ***Α*** που υπάρχει στο τέλος της εργασίας παρουσιάζονται τα ελαττώματα τα οποία αφορούν τα εξαρτήματα και τα οποία είναι τοποθετημένα στις πλακέτες του προϊόντος. Οι σταθμοί ελέγχου ποιότητας είναι υπεύθυνοι για την ανίχνευση ύπαρξης κάποιου ελαττώματος. Για την ευκολία των ελέγχων και την καλύτερη επικοινωνία των διεργασιών και των εργαζομένων, η εταιρεία έχει κωδικοποιήσει τα ελαττώματα και έχει φτιάξει λίστες οι οποίες περιέχουν όλες τις κατηγορίες των ελαττωμάτων. Οι κωδικοί των ελαττωμάτων και η επεξήγηση τους παρουσιάζονται στον πίνακα Ι

Στον ***πίνακα Ι*** παρατίθενται οι κωδικοί ελαττωμάτων όπως και η επεξήγηση τους. Στην πρώτη στήλη αναφέρεται ο σταθμός ελέγχου που ανιχνεύει τα ελαττώματα της δεύτερης στήλης. Οι κωδικοί ελαττωμάτων της δεύτερης στήλης είναι τα χαρακτηριστικά ποιότητας που εξετάζονται από τον αντίστοιχο σταθμό ελέγχου. Στην τέταρτη στήλη αναφέρονται οι σταθμοί οι οποίοι χρεώνονται τα ελαττώματα που ανιχνεύονται στη δεύτερη στήλη. Προφανώς στην πρώτη περίπτωση πηγή χρέωσης είναι μόνο ο σταθμός P1 διότι μόνο αυτός υπάρχει ως προηγούμενος του σταθμού ελέγχου του. Ως FM χαρακτηρίζεται το ελάττωμα που προέρχεται από εξωτερική πηγή (π.χ. προμηθευτής).

Κεφάλαιο 2^0

Bayesian Networks

Αυτό το κεφάλαιο αποτελεί μια σύντομη εισαγωγή στα Bayesian Networks.Αφού οριστεί το μοντέλο Μπααεζιανών δικτύων,επικεντρωνόμαστε στις σημαντικές ιδιότητες για την δομική εκπαίδευση (structural learning).

2.1 Εισαγωγή

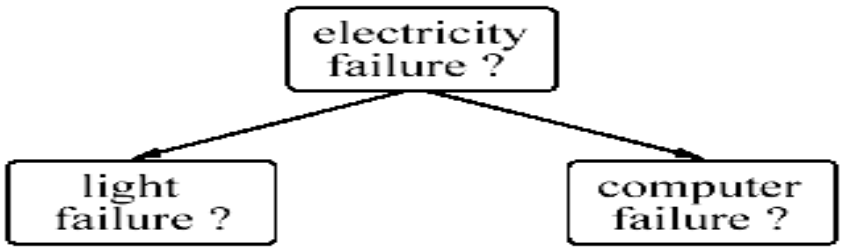
Η αβεβαιότητα είναι παρούσα σε πολλές δραστηριότητες στη ζωή μας,εμποδίζοντας μας,στις αποφάσεις που παίρνουμε.Για παράδειγμα ,στην ιατρική μερικές φορές είναι δύσκολο να καταλήξουμε στην πάθηση ενός ασθενούς, βασιζόμενοι στα παρατηρούμενα συμπτώματα.Αυτό συμβαίνει γιατί εμπεριέχει ένα συγκεκριμένο σύμπτωμα με κάποια πιθανότητα ,αλλά χωρίς βεβαιότητα. Επιπλέον μερικά συμπτώματα μπορούν να χαρακτηριστούν ως «χαλαρά»(π.χ υψηλός πυρετός δεν ξεκινάει σε κάποια συγκεκριμένη θερμοκρασία)Επίσης,η αρρώστεια στην οποία καταλήξαμε βασίζεται σε ανολοκλήρωτη γνώση,γιατί ένας γιατρός μπορεί να μη γνωρίζει ,για παράδειγμα , τη πληροφόρηση σχετικά με την επίσκεψη του ασθενούς, σε μια ορισμένη χώρα,παράγοντας που αυξάνει τις πιθανότητες μιας μόλυνσης σε ένα ιδιαίτερο είδος βακτηριδίων. Σε πολλούς τομείς των επιστημών και της εφαρμοσμένης μηχανικής, όταν μια ακριβής περιγραφή ενός μεγάλου συστήματος περιλαμβάνει πάρα πολλά στοιχεία και δεδομένα, τότε συχνά προσφεύγουν σε ένα *προσεγγιστικά* μοντέλα που είναι μια καλή απόδοση του πραγματικού συστήματος. Αυτό είναι μια άλλη πηγή αβεβαιότητας. Παραδείγματος χάριν, στην τεχνητή νοημοσύνη η ανάλυση των κειμένων είναι συχνά βασισμένη σε μια πιθανολογική επεξεργασία, όπου το κείμενο θεωρείται ως "*bag of words* ", δηλ.λέξες χωρίς λογική σύνταξη.Τα αναπτυσόμενα ευφυή συστήματα που λειτουργούν σε περιβάλλον αβεβαιότητας είναι μια από τις κύριες προκλήσεις στον τομέα της τεχνητής νοημοσύνης (AI). Διάφορες προσπάθειες έχουν καταβληθεί για την αντιμετώπιση της αβεβαιότητας στον τομέα της AI, όπως την « ασαφή» λογική (**fuzzy logic**) ή την θεωρία **Dempster- Schafer** . Κατά τη διάρκεια των προηγούμενων μιας ή δύο δεκαετιών, η θεωρία πιθανότητας έχει αυξανόμενη επίδραση στην κοινότητα AI, δεδομένου ότι είναι μια «υγιής» θεωρία για την εξέταση της αβεβαιότητας. Η θεωρία πιθανότητας έχει εφαρμοστεί στις στατιστικές προκειμένου να αντληθούν πληροφορίες από τα δεδομένα. Αυτό γίνεται χαρακτηριστικά με τη βοήθεια των αποκαλούμενων δοκιμών υπόθεσης(hypothesis tests), όπου μια υπόθεση όπως *:" το κάπνισμα ασκεί επίδραση στον καρκίνο πνευμόνων;"* ικανοποιείται ή όχι.

Για αρκετό χρονικό διάστημα, η στατιστική ανάλυση είχε περιοριστεί μόνο σε έναν *μικρό* αριθμό μεταβλητών. Αυτό είχε δύο λόγους. Κατ' αρχάς, η στατιστική ανάλυση των σύνθετων υποθέσεων που περιλαμβάνουν έναν *μεγάλο* αριθμό μεταβλητών δεν είχε κατανοηθεί πλήρως ,από τη θεωρητική σκοπιά.. Δεύτερον, η ανάλυση των περίπλοκων υποθέσεων ήταν δυσεπίλυτη έως ότου οι ισχυροί

υπολογιστές διατέθηκαν. Τα Μπαεζιανά δίκτυα έχουν γίνει το σταθερό μοντέλο για την αντιμετώπιση της αβεβαιότητας στην AI. Αναπτύχθηκαν από την κοινότητα της AI για να χτίσουν τα πιθανολογικά έμπειρα συστήματα τα οποία λειτουργούν κάτω από την αβεβαιότητα. Ένα από τα κύρια πλεονεκτήματά τους είναι η υγιής θεωρητική βάση τους στα πλαίσια των στατιστικών και της θεωρίας πιθανότητας. Ένα Μπεϋζιανό δίκτυο είναι ένα *πιθανολογικό μοντέλο* περιγραφής της *πολλαπλή-μεταβλητών(multivariate)* κατανομή πιθανότητας για ένα σύνολο μεταβλητών.

Ειδικότερα, σχεδιάζεται για τις περιοχές (domains) με έναν *μεγάλο* αριθμό μεταβλητών.Η βασική ιδέα είναι να επιδειχθούν οι συσχετισμοί μεταξύ των μεταβλητών, δηλαδή οι κατά συνθήκη*(conditional) ανεξαρτησίες και οι εξαρτήσεις*, με τη βοήθεια μιας γραφικής παράστασης. Για αυτόν τον λόγο, τα Μπεϋζιανά δίκτυα ανήκουν στην κατηγορία *γραφικών μοντέλων(**graphical models**)*. Το σχήμα 2.1 παρουσιάζει απλοϊκή Μπαεζιανή δομή δικτύων, όπου οι άκρες αντιπροσωπεύουν μια εξάρτηση μεταξύ των δυαδικών(binary) μεταβλητών *"πτώση ηλεκτρικής ενέργειας;"* και *" έλλειψη φωτισμού;"* όπως και μεταξύ των πρώτων και της *"μη λειτουργία του υπολογιστή;"*. Η απουσία τόξου μεταξύ της μεταβλητής *"έλλειψη φωτισμού;"* και *" μη λειτουργία του υπολογιστή;"*, δείχνει ότι αυτές οι δύο μεταβλητές είναι ανεξάρτητες υπό όρους στη "πτώση ηλεκτρικής ενέργειας;". Αυτό σημαίνει ότι η μεταβλητή *" έλλειψη φωτισμού;"* είναι *ασυσχέτιστη* με την μεταβλητή *"μη λειτουργία του υπολογιστή*;", και αντίστροφα,όταν η κατάσταση της *" πτώση ηλεκτρικής ενέργειας;"* είναι γνωστή. Παραδείγματος χάριν, εάν η μια γνωρίζει ότι δεν υπάρχει καμία αποτυχία ηλεκτρικής ενέργειας, η λάμπα φωτισμού και ο υπολογιστής αποτυγχάνουν ανεξάρτητα ο ένας από τον άλλον, όπως κάθε ένας από τους οποίους μπορεί να υποστεί ζημιά κατά τύχη. Εντούτοις, όταν η κατάσταση της *" πτώση ηλεκτρικής ενέργειας*;" είναι άγνωστη, τότε η *"έλλειψη φωτισμού*;"και *" μη λειτουργία υπολογιστών;"* εξαρτάται η μια από την άλλη, όπως και οι δύο εξαρτώνται από τη *"πτώση ηλεκτρικής ενέργειας;"* στο σχήμα 2.1.

Αυτό είναι επίσης διαισθητικά σαφές, δεδομένου ότι μια αποτυχία ηλεκτρικής ενέργειας συνεπάγεται και την πτώση του φωτισμού όπως και τις αποτυχία/πτώση τάσης του υπολογιστή. Είναι ως εκ τούτου κρίσιμο να γίνει διάκριση μεταξύ των "άμεσων" και "έμμεσων" ενώσεων μεταξύ των μεταβλητών .



Σχήμα 2.1: Μια απλοποιημένη δομή ενός Μπαεζιανού δικτύου που δείχνει τις άμμεσες(direct) και των έμμεσων συσχετίσεων (indirect) μεταξύ αυτών των τριών μεταβλητών.

Η γραφική αναπαράσταση των σχέσων των μεταβλητών του σχήματος 2.1 μπορεί επίσης να γίνει κατανοητή διαισθητικά όταν ερμηνεύεται κατά τρόπο

αιτιώδη, δηλαδή "πτώση ηλεκτρικής ενέργειας;" Μπορεί να θεωρηθεί ως η κοινή αιτία της " έλλειψη του φωτισμού;" και "έλλειψη του υπολογιστή;". Αυτό υπονοεί ότι το περιστατικό της " έλλειψη του φωτισμού;" και της "μη λειτουργία του υπολογιστή;" μπορεί να συσχετιστούν όταν δεν είναι τίποτα γνωστό για τη "πτώση ηλεκτρικής ενέργειας;".Αντιθέτως, όταν είναι γνωστή η κατάσταση της κοινής αιτίας, η μεταβλητή"έλλειψη φωτισμού ;" και "μη λειτουργία του υπολογιστή;" μπορούν να θεωρηθούν ανεξάρτητες. Όταν τα Μπεϋζιανά δίκτυα χρησιμοποιούνται σε μια αιτιώδη ρύθμιση (causal settings), ονομάζονται επίσης και *αιτιώδη δίκτυα (**causal networks**)*. Η χρήση τους ως αιτιώδη δίκτυα είναι πολύ ελκυστική στην κοινότητα AI δεδομένου ότι η γνώση σχετικά με τις αιτιώδεις σχέσεις καθιστά τις καλά κατανοημένες παρεμβάσεις εφικτές και ως εκ τούτου επιτρέπει σε κάποιον να έχει τον έλεγχο της συμπεριφορά των σύνθετων συστημάτων. Η αιτιώδης ερμηνεία μιας Μπεϋζιανής δομής δικτύων είναι, εντούτοις, μόνο κατάλληλη υπό ορισμένους όρους/συνθήκες. Εκτός από την απεικόνιση των υπό όρους/κατα συνθήκη ανεξαρτησιών και των εξαρτήσεων, ένα μοντέλο Μπαεζιανό δικτύων μπορεί επίσης να περιγράψει την κοινή κατανομή πιθανότητας(***joint probability distribution***) για ένα σύνολο μεταβλητών κατά τρόπο *ποσοτικό*, δεδομένου ότι ένα Μπαεζιανό δίκτυο περιλαμβάνει επίσης τις παραμέτρους εκτός από τη γραφική δομή του. Είναι ευεργετικό ότι αυτή η περιγραφή είναι αρκετά τροπική(modular) έτσι ώστε τα σύνθετα συστήματα μπορούν να χαρακτηριστούν από το συνδυασμό των μικρότερων μονάδων. Όταν οι τιμές μερικών μεταβλητών είναι γνωστές, ένα Μπαεζιανό δίκτυο μπορεί να χρησιμοποιηθεί για την πρόβλεψη των καταστάσεων των άλλων μεταβλητών.

Αυτή η διαδικασία καλείται *συμπέρασμα(**inference**)*. Το αποτέλεσμα δεν είναι μια συγκεκριμένη τιμή/κατάσταση , αλλά μια ***κατανομή πιθανότητας*** που αναφέρεται στις καταστάσεις των μεταβλητών οι οποίες προβλέπονται, δηλ. ένα Μπεϋζιανό δίκτυο περιγράφει *τους πιθανολογικούς* συσχετισμούς μεταξύ των μεταβλητών σε μια περιοχή***(domain***). Δεδομένου ότι ο αναγνώστης να εξοικειωθεί με τα νευρικά δίκτυα, αναφέρετε ότι, παραδείγματος χάριν στα ***feed-forward*** νευρικά-δίκτυα, η κατάσταση μιας μεταβλητής εξόδου εξαρτάται ***deterministically/με βεβαιότητα*** από τις μεταβλητές εισόδου, παρά πιθανοτικά/probabilistically. Υπάρχουν βασικά δύο τρόποι δημιουργίας ενός Μπαεζιανού δίκτυο.

Κατ' αρχάς, ένα σύστημα που γίνεται κάλα-κατανοητό μπορεί να διαμορφωθεί με το χέρι από τους εμπειρογνώμονες. Μια αιτιώδης ερμηνεία των τόξων και των προσανατολισμών τους είναι συχνά χρήσιμη σε αυτήν την κατάσταση. Το αποκαλούμενο "*Μπαεζιανό δίκτυο προσανατολισμένου αντικειμένου*" ***(object oriented Bayesian Networks)*** αναπτύχθηκε για να βοηθήσει την κατασκευή των μεγάλων μοντέλων. Σημαντικοί τομείς εφαρμογής τους ήταν συστήματα για ιατρική διάγνωση και για την ανίχνευση λαθών. Εναλλακτικά, τα Μπαεζιανά δίκτυα μπορούν *να εκπαιδευτούν(**learning**)* από τα δεδομένα . Αυτό σημαίνει ότι ένας αλγόριθμος εκμάθησης μπορεί να προκαλέσει –δημιουργήσει τη δομή καθώς επίσης και τις παραμέτρους ενός Μπαεζιανού δικτύου από τα δεδομένα. Φυσικά, επίσης ένας συνδυασμός και των δύο προσεγγίσεων είναι δυνατός, όπου η γνώση του ειδικού μπορεί να χρησιμεύσει ως η προγενέστερη γνώση που ενσωματώνεται σε έναν αλγόριθμο εκμάθησης. Στη δομική εκπαίδευση(learning structure) των Μπαεζιανών δικτύων, όλες οι μεταβλητές σε μια περιοχή/domain αντιμετωπίζονται ισοδύναμα, δηλ. δεν υπάρχει καμία ευδιάκριτη κατηγορία-μεταβλητής(distinct class variable).Αυτό είναι χαρακτηριστικό για *την ανεπίβλεπτη* εκπαίδευση(***unsupervised learning***).

Δεδομένου ότι μια δημιουργημένη Μπαεζιανή δομή δικτύων απεικονίζει τους συσχετισμούς μεταξύ των διάφορων μεταβλητών σε μια περιοχή, η εκπαίδευση/εκμάθηση της δομής μπορεί να παρέχει τις νέες ιδέες που βοήθουν στην κατανόηση των συσχετισμών μεταξύ των μεταβλητών. Η εξαγωγή της νέας γνώσης από τα δεδομένα καλείται *ανάσυρση δεδομένων(**data mining**)*. Οι αυτόματες διαδικασίες εκμάθησης μπορούν ως εκ τούτου να συμπληρώσουν-βελτιώσουν την στατιστική ανάλυση των δεδομένων. Η εκτίμηση των παραμέτρων ενός Μπαεζιανού δικτύου είναι ένα χαρακτηριστικό υποπρόβλημα της εκμάθησης της δομής του.Αφότου ένα Μπαεζιανό δίκτυο, συμπεριλαμβανομένων των παραμέτρων του, έχει εκπαιδευτεί, μπορεί να χρησιμοποιηθεί για τις ποσοτικές προβλέψεις.

Το συνεργάσιμο φιλτράρισμα(***collaborative filtering***) είναι μια πρόσφατη εφαρμογή όπου τα Μπεϋζιανά δίκτυα έχουν αποδείξει ότι παραγάγουν πολύ ακριβή αποτελέσματα. Αυτό υιοθετείται, παραδείγματος χάριν, στα απευθείας σύνδεση βιβλιοπωλεία(on line bookstores) όπου τα Μπαεζιανά δίκτυα εκπαιδεύονται στα δεδομένα που συλλέγονται από τους πελάτες, και που χρησιμοποιούνται στη συνέχεια για να συστήσουν εκείνοι τα βιβλία που έχουν μια υψηλή πιθανότητα της ύπαρξης ενδιαφέροντος από έναν πελάτη. Το *World Wide Web* θα είναι βεβαίως μια κύριως χώρος για την ανάσυρση δεδομένων, καθώς επιτρέπει την πρόσβαση σε ένα τεράστιο ποσό στοιχείων, ενώ οι δαπάνες για τη συλλογή στοιχείων μειώνονται παρά πολύ συγχρόνως. Η βιοτεχνολογία και η γενετική είναι ένας άλλος τομέας όπου οι τεχνικές ανάσυρσης δεδομένων μπορούν να βοηθήσουν στην κατανόηση των διαδικασιών που κρύβονται κάτω από τα στοιχεία.

2.2 Η βασισμένη σε περιορισμούς προσέγγιση εφαρμοσμένη σε πεπερασμένα σύνολα δεδομένων (The Constraint-Based Approach applied to Finite Data Sets)

Η αποτελεσματικότητα της προσέγγισης βασισμένης σε περιορισμούς προκαλεί την αραίωση των δομών των Μπεϋζιανών δικτύων και είναι πολύ ελκυστική ως μέθοδος. Οι άλλοι ευρετικοί αλγόριθμοι εκμάθησης απαιτούν αρκετό χρόνο, ιδιαίτερα στις περιοχές με έναν μεγάλο αριθμό μεταβλητών. Αντίθετα από άλλους αλγορίθμους εκμάθησης, η βασισμένη σε περιορισμούς προσέγγιση απαιτεί δύο πρόσθετες υποθέσεις για τη κατανομή πιθανότητας που υπονοείται από τα στοιχεία. Ας αναφέρουμε αυτές τις δύο υποθέσεις

- η κατανομή πιθανότητας είναι πλήρως γνωστή, δηλ. χωρίς λάθος, και
-η κατανομή πιθανότητας εκπληρεί την αποκαλούμενη *υπόθεση πίστης(**faithfulness assumption**).*

Εάν αυτές οι υποθέσεις ισχύουν, η βασισμένη σε περιορισμούς προσέγγιση μπορεί να αποδειχθεί ότι να παραγάγει τη *σωστή* δομή Bayesian δικτύων ,ο αποκαλούμενος τέλειος χάρτης(***perfect map***). Σημειώστε ότι αυτό δεν είναι εγγυημένο για τις άλλες προσεγγίσεις που στοχεύουν στη βελτιστοποίηση μιας λειτουργίας αποτελέσματος(***scoring function***), καθώς μπορούν να παραμείνουν κολλημένοι στα τοπικά βέλτιστα.

Στην πράξη, εντούτοις, αυτές οι υποθέσεις δεν χρειάζονται να ισχύουν. Στην πραγματικότητα, μπορούν μόνο να αναμένονται για να κρατήσουν μέσα στο ασυμπτωτικό όριο, δηλ. όταν ένα άπειρο ποσό στοιχείων είναι διαθέσιμο. Αυτό συμβαίνει επειδή ο θόρυβος δειγματοληψίας είναι χαρακτηριστικά παρών *στα πεπερασμένα* σύνολα δεδομένων, που αναγκάζουν τη κατανομή πιθανότητας που υπονοείται/εμφανίζεται από τα στοιχεία να διαφέρει από την *αληθινή* κατανομή

(από την οποία τα στοιχεία είχαν επιλεχθεί).Αυτό υπονοεί ότι η βασισμένη σε περιορισμούς προσέγγιση παράγει γενικά "σχεδόν τις σωστές " δομές Bayesian δικτύων δεδομένου των αρκετά μεγάλων συνόλων δεδομένων ,όπως επιβεβαιώνεται από πολλά πειράματα που αμναφέρονται στη βιβλιογραφία. Εκτός από αυτά τα πειράματα, πολλή προσοχή δεν έχει δοθεί στη συμπεριφορά της βασισμένης σε περιορισμούς προσέγγισης όταν τα *πεπερασμένα* σύνολα στοιχείων είναι γνωστά. Τα πεπερασμένα παρά τα άπειρα σύνολα δεδομένων είναι, εντούτοις, χαρακτηριστικά για τις πρακτικές εφαρμογές. Λαμβάνοντας υπόψη *τα πεπερασμένα* μεγέθη των δειγμάτων, οι δύο ανωτέρω υποθέσεις δεν είναι εγγυημένες να εκπληρωθούν. Αυτό είναι ιδιαίτερα προφανές σχετικά με την πρώτη υπόθεση, δεδομένου ότι είναι γνωστό στις στατιστικές ότι μια δοκιμή ανεξαρτησίας μπορεί να αποτύχει, δηλ. μια εξάρτηση μπορεί λανθασμένα να προκληθεί αντί μιας ανεξαρτησίας, και αντίστροφα. Αυτό καλείται ***σφάλμα τύπου I*** και ***σφάλμα τύπου II*** , αντίστοιχα.

Τα αποτελέσματα *διάφορων* δοκιμών συνδυάζονται με τη βασισμένη σε περιορισμούς προσέγγιση προκειμένου να κατασκευαστεί η δομή των Μπαεζιανών δικτύων. Δεδομένου ότι μερικά από τα αποτελέσματα της δοκιμής να είναι ανακριβή και δεδομένου ότι τα αποτελέσματα διαφόρων δοκιμών(tests) να εξαρτώνται το ένα από το άλλο με κάποιο άγνωστο τρόπο, το λάθος της δημιουργημένης δομής των Μπαεζιανών δικτύων δεν είναι υπό έλεγχο. Στις στατιστικές, αυτό είναι ένα γνωστό πρόβλημα στον τομέα της πολλαπλάσιας δοκιμής. Ας αναφέρουμε τη βασική ιδέα στα εξής: τις ανωτέρω υποθέσεις που κρύβονται κάτω από τους αλγορίθμους βασισμένους σε περιορισμούς εγκαταλείπονται. Αντ' αυτού, προτιμάμε την άποψη ότι η βασισμένη σε περιορισμούς προσέγγιση στοχεύει στην εύρεση της βέλτιστης δομής των Μπαεζιανών δικτύων όσον αφορά μια συνάρτηση αποτελέσματος(scoring function). Η χρήση μιας συνάρτησης αποτελέσματος είναι παρόμοια με την εναλλακτική προσέγγιση στη δομική εκμάθηση .Η άποψη αυτή ενισχύεται από το γεγονός ότι η χρήση μιας συνάρτησης αποτελέσματος γίνεται καλά κατανοητή, και η βέλτιστη δομή των Μπαεζιανων δικτύων είναι καθορισμένη με σαφήνεια επίσης σε εκείνες τις περιπτώσεις όπου *τα πεπερασμένα* σύνολα στοιχείων δίνονται

Προκειμένου να γίνει κατανοητή η βασισμένη σε περιορισμούς προσέγγιση στα πλαίσια της βελτιστοποίησης μιας συνάρτησης αποτελέσματος , εισάγεται η έννοια *των σχετικών* συναρτήσεων αποτελέσματος.Ενδιαφέρονται για τις διαφορές των αποτελεσμάτων παρά για τα αποτελέσματα των ιδίων.. Αυτό επιτρέπει σε μας να κάνουμε χρήση των συναρτήσεων αποτελέσματος όπως το Μπαεζιανό κριτήριο πληροφοριών ή τη μεταγενέστερη πιθανότητα σε αυτήν την προσέγγιση, αντί x^2 - δοκιμή που υιοθετείται συνήθως από τους αλγορίθμους βασισμένους σε περιορισμούς. Η άποψή μας αποκαλύπτει ότι οι δομές δικτύων που προκαλούνται από την βασισμένη σε περιορισμούς προσέγγιση μέσω των πεπερασμένων συνόλων στοιχείων που τείνουν να περιέχουν πολύ λίγες ακμές, έναντι της βέλτιστης γραφικής παράστασης.

Επιπλέον, η απόδοση της βασισμένης σε περιορισμούς προσέγγιση μπορεί να βελτιωθεί αρκετά με τη χρησιμοποίηση της αποκαλούμενης *απαραίτητης συνθήκης πορείας* ***(necessary path condition)*** απόν την οποία αντλούμε από τις ιδιότητες των βέλτιστων δομών των Μπαεζιανών δικτύων και στην οποία θα αναφερθούμε εκτενέστερα στο τέταρτο κεφάλαιο**.**Επιπλέον, η *αβεβαιότητα του μοντέλου* μπορεί να ανακαλυφθεί κατά χρησιμοποίηση αυτής της επέκτασης. Φυσικά, μπορεί μόνο να εξερευνηθεί μέχρι κάποιο βαθμό, δεδομένου ότι μια ακριβής επεξεργασία αυτής της αβεβαιότητας είναι ανέφικτη εκτός από τις περιοχές με έναν μάλλον μικρό αριθμό μεταβλητών. Η αβεβαιότητα του μοντέλου επικρατεί γενικά όταν δίνονται τα μικρά

σύνολα στοιχείων. Επειδή τα σύνολα δεδομένων, ακόμα και όταν θεωρούνται ως "μεγάλα",μπορεί συχνά να είναι μικρά έναντι του αριθμού κοινών καταστάσεων των μεταβλητών σε μια μεγάλη περιοχή, η αβεβαιότητα του μοντέλου θα ήταν καλύτερο να μην αγνοηθεί σε πολλές εφαρμογές .

Υπολογίζοντας την αβεβαιότητα του μοντέλου μπορεί μ'αυτό τον τρόπο να βελτιωθεί η προβλεπόμενη/προφητική(*predictive*) ακρίβεια ,π.χ. η αποφυγή των ανακριβών συμπερασμάτων που προέρχονται από τις κατασκευασμένες δομές βοηθάει στην ακρίβεια του μοντέλου. Σχετικά με το τελευταίο ζήτημα, δείχνουμε ότι οι πολλαπλάσιες λύσεις που προκύπτουν με τη βοήθεια της απαραίτητης συνθήκης πορείας μπορούν να απεικονιστούν σε μια ενιαία γραφική παράσταση. Αυτό συμβαίνει επειδή οι διάφορες γραφικές παραστάσεις έχουν συνήθως πολλές άκρες από κοινού και διαφέρουν μόνο σχετικά με την παρουσία μερικών ακμών. Μια τέτοια γραφική παράσταση μπορεί χαρακτηριστικά να ερμηνευθεί ευκολότερα από έναν κατάλογο των διάφορων λύσεων. Η απαραίτητη συνθήκη πορείας μπορεί αποτελεσματικά να εφαρμοστεί από την άποψη των κανόνων. Η παρουσία ακμών προκαλείται με την απλούστευση του συνόλου κανόνων. Αυτό καθιστά μια συστηματική κατασκευή όλων των πολλαπλάσιων λύσεων πιθανά ,τα οποία μπορούν να προκληθούν με τη βοήθεια της απαραίτητης συνθήκης πορείας. Επιπλέον, αυτό το σχέδιο μπορεί να εκμεταλλευθεί το γεγονός ότι οι διάφορες δομές των Μπαεζιανών δικτύων έχουν χαρακτηριστικά πολλές άκμες από κοινού. Για αυτόν τον λόγο, το υπολογιστικό κόστος των προκαλούμενων ενδεχόμενων γραφημάτων αυξάνεται ελαφρώς έναντι των καθιερωμένων βασισμένων στον περιορισμό προσεγγίσεις βασισμένες στους περιορισμούς, οι οποίες προκαλούν μόνο μια ενιαία γραφική παράσταση εξ ορισμού. Η προσέγγιση βασισμένη σε περιορισμούς είναι επίσης κατάλληλη για τον παράλληλο υπολογισμό που έχουμε πραγματοποιήσει σε ένα απλό σχέδιο κυρίου(**master**) και σκλάβων(**slave scheme**).

Σε περίπτωση που η κατανομή πιθανότητας που υπονοείται από τα στοιχεία εκπληρώνει τις υποθέσεις που απαιτούνται από την κατάσταση προόδου των προσεγγίσεων που βασίζονται σε περιορισμούς, η απαραίτητη συνθήκη πορείας παράγει την ίδια γραφική παράσταση όπως οι άλλες προσεγγίσεις βασισμένες σε περιορισμούς, και είναι ως εκ τούτου ασυμπτωτικά σωστό. Λαμβάνοντας υπόψη τα πεπερασμένα στοιχεία, εντούτοις, η συνθήκη απαραίτητης πορείας είναι μια σημαντική επέκταση της προσέγγισης βασισμένης στον περιορισμό.

Επίσης υποστηρίζουμε ότι οι άκρες που προκαλούνται για να είναι παρούσες από την προσέγγιση βασισμένη σε περιορισμούς περιλαμβάνονται σε μια (τοπικά) βέλτιστη δομή των Μπαεζιανών δικτύων με έναν υψηλό βαθμό βεβαιότητας.Αυτό δείχνει την αναγκαιότητα ενός βήματος εκπαίδευσης ,επακόλουθο της προσέγγισης βασισμένη σε περιορισμούς,εάν κάποιος στοχεύει στη πρόκληση(τοπικά) των βέλτιστων δομή των Μπαεζιανών δικτύων.Πρωτού προχωρήσουμε σε περαιτέρω ανάλυση της θεωρίας των Bayesian Networks πρέπει να σημειώσουμε ότι στο παράρτημα ***B*** παρουσιάζονται οι κυριότεροι συμβολισμοί για την καλύτερη κατανοήση εξισώσεων και των εννοιών που υπάρχουν παρακάτω.

2.3 **Κατά συνθήκη ανεξαρτησίες και εξαρτήσεις** *(conditional independences and dependences)*

Ένα Μπαεζιανό δίκτυο είναι ένα πιθανολογικό μοντέλο βασισμένο στην έννοια ***των υπό όρους ανεξαρτησιών και των εξαρτήσεων (CIDs).*** Ως εκ τούτου θα ήταν φρόνιμο κάποια εισαγωγικό σημείωμα από τη θεωρία πιθανότητας Η κατανομή πιθανότητας για μια τυχαία μεταβλητή a δηλώνεται ως ***p(a),*** και η *κοινή* πιθανότητα για διάφορες μεταβλητές, π.χ. για ***S ={a,b,c}*** ,τις αναγνωρίζει ως ***p(a, b, c)*** ή ***p (S)***. Περιγράφει την πιθανότητα για κάθε *κοινή κατάσταση* ή *διαμόρφωση* των μεταβλητών στο καθορισμένο **S.** Αφήστε το σύνολο όλων των διαμορφώσεων(configurations) των μεταβλητών στο καθορισμένο **S** να συμβολιστεί **I(S),** και οι διαμορφώσεις ως ***ι, j, K... Є I(S*).** Οι ανωτέρω πιθανότητες καλούνται επίσης ***οριακές* πιθανότητες(marginal probabilities)**, σε αντιδιαστολή με ***τις κατά συνθήκη* πιθανότητες(conditional probabilities)**. Η πιθανότητα μιας μεταβλητής ***a*** υπό όρους σε ένα σύνολο μεταβλητών, π.χ. ***S = {b,c}* , υποδεικνύεται ως *p(a/b, c)*** ή **p(a/S),** και ορίζεται ως ***p(a/b, c) = p(a, β, c)=p(b, c)*** για ***p(b, c) > 0***. Δύο τυχαίες μεταβλητές a και b θεωρούνται ***οριακά ανεξάρτητες(marginally independent)***όταν για την κοινή πιθανότητά τους ***p(a, b)***ισχύει ότι : ***p(a, b) = p (a)p (b)***. Μια τέτοια ανεξαρτησία δείχνεται δεδομένου ότι $a \perp\!\!\!\perp b$ σημαίνει ότι η κατάσταση του a είναι ασυσχέτιστη με την κατάσταση του b, και αντίστροφα. Ομοίως, δύο μεταβλητές a και b είναι *ανεξάρτητες υπό όρους* σε μερικές άλλες μεταβλητές που περιλαμβάνονται στο καθορισμένο S εάν οι κατά συνθήκη πιθανότητες όπως το ***p (α, b/S) = p (a/S)p (b/S)*** δεδομένο ***p (S) > 0***. Αυτό είναι ισοδύναμο με το***p(a/b, s) = p (a/S) ή p(b/a, S)= p (b/S)*** υπό τον όρο ότι ***p(a, S), p(b/S) ,p (S) > 0***, Αυτό σημαίνει ότι το b είναι άσχετο με εάν η κοινή κατάσταση των μεταβλητών στο καθορισμένο s είναι γνωστή. Μια τέτοια υπό όρους ανεξαρτησία μπορεί να δειχτεί ως $a \perp\!\!\!\perp b|s$

Εάν δύο μεταβλητές δεν είναι ανεξάρτητες αυτόματα θεωρούνται *εξαρτημένες*. Αν κάποιος μπορεί να διακρίνει μεταξύ μιας οριακής εξάρτησης δύο μεταβλητών α και b, που υποδεικνύεται ως $a \not\perp\!\!\!\perp b$ και μια εξάρτηση υπό όρους σε καθορισμένο δείγμα ως: $a \not\perp\!\!\!\perp b|S$

2.4 Μπαεζιανά δίκτυα και οι ιδιότητές τους

Σε όλη αυτήν την διατριβή, αφήστε το σύνολο *όλων* των μεταβλητών σε μια περιοχή να δειχτούν ως ***V*** , και τις μεταβλητές ως ***a,b... Є V*** . Οι ***υπό όρους ανεξαρτησίες και οι εξαρτήσεις* (CIDs)** που κρύβονται κάτω από μια πολλαπλών μεταβλητών κατανομή πιθανότητας για τις μεταβλητές στο V απεικονίζονται από τη *γραφική δομή* ενός Μπαεζιανού δικτύου, η αποκαλούμενη *κατευθυνόμενη ακυκλική γραφική παράσταση/directed acyclic graph* (**DAG**). Περιγράφεται στο επόμενο τμήμα. Το άλλο συστατικό ενός Μπαεζιανού δικτύου είναι ένα σύνολο *παραμέτρων*, που καθιστούν εφικτή μια ποσοτική περιγραφή των διανομών πιθανότητας

2.4.1 Directed Acyclic Graph (DAG)

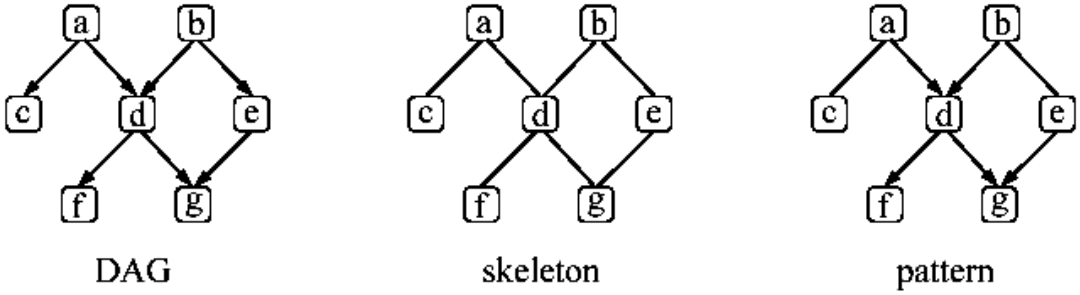
Αρχικά ενδιαφερόμαστε για την κατευθυνόμενη ακυκλική γραφική παράσταση (***DAG) m,***δηλαδή για τη δομή των Μπαεζιανών δικτύων Συσχετίζεται με τις κατανομές πιθανότητας κατά τέτοιο τρόπο ώστε οι τυχαίες μεταβλητές αντιστοιχούν στους κόμβους ή τόξα στο ***DAG***. Εκτός από τις μεταβλητές, *οι κατευθυνόμενες* άκμες(directed edges) και τα τόξα (vertices) είναι παρούσα σε ένα ***DAG***. Καμία απροσανατολισμένη άκρη δεν επιτρέπεται.Σ ' αυτή την εργασία θα γίνει χρήση τόσο των μεταβλητών ,όσο των κόμβων όσο και των τόξων.

Μία κατευθυνόμενη ακμή η οποία έχει φορά από την μεταβλητή a προς την μεταβλητή b συμβολίζεται ως:$\alpha \rightarrow b$ ***ή*** $b \leftarrow \alpha$ Όταν οι κατευθύνσεις των ακμών αγνοούνται ένα απροσανατολισμένο γράφημα δημιουργείται ,το οποίο καλείται σκελετός(skeleton) του DAG m. Έστω ότι ο σκελετός συμβολίζεται με $\bar{m}$.Μία απροσανατολισμένη ακμή μεταξύ δύο μεταβλητών $a, b \in V$ συμβολίζεται ως $\alpha \sim b$. Ένα μονοπάτι αν,αμεσα σε δύο μεταβλητές a, b $\in$ V είναι μια αλληλουχία ακμών $a=x_0 \sim x_1 \sim \ldots \sim x_r=b$ ανεξάρτητα από τις κατευθύνσεις τους ένα κατευθυνόμενο μονοπάτι από την ***a*** στη ***b*** αποτελείται από μια ακολουθία ακμών

a $=x_0 \rightarrow x_1 \rightarrow \ldots \rightarrow x_r=b$ έτσι ώστε: $x_{i-1} \rightarrow x_i$ για όλα τα $i=1,\ldots,r$.Είναι ζωτικής σημασίας ότι η δομή των Μπαεζιανών δικτύων και επομένως η έννοια μιας κατευθυνόμενης *ακυκλικής* γραφικής παράστασης δεν περιέχει *τους κατευθυνόμενους κύκλους*, δηλ. μια κατευθυνόμενη πορεία που αρχίζει και τελειώνει στην ίδια μεταβλητή. Αυτό είναι διευκρινισμένο στο απλοϊκό ***DAG*** που παρουσιάζεται στο σχήμα 2.2, όπου κανένας κατευθυνόμενος κύκλος δεν εμφανίζεται αν και οι μεταβλητές ***b,d, g*** και ***e*** περιλαμβάνονται σε έναν *βρόχο*, δηλ. ο αντίστοιχος σκελετός περιέχει μια κλειστή πορεία.

Η ακόλουθη παρατήρηση εμπνέεται από ένα οικογενειακό δέντρο. Οι *γονείς* $pa_m(a)$ μιας μεταβλητής ***a*** στο **DAG m** είναι το σύνολο μεταβλητών $v \in V$ έτσι ώστε υπάρχει μια κατευθυνόμενη ακμή $v \rightarrow a$. Η μεταβλητή ***a*** καλείται παιδί της ***υ*** .Αν οι προσανατολισμοί παραβλέπονται,η έννοια των «γειτόνων» μιας μεταβλητής $a \in V$ είναι χρήσιμη και συμβολίζεται ως ***ne(a***). Το σύνολο ***ne(a)*** περιέχει όλες τις μεταβλητές που είναι γειτονικές της μεταβλητής **α**.Όταν διάφορες "γενεές" εξετάζονται και μια μεταβλητή $v \in V$ είναι πρόγονος(***ancestors***) της ***α***, σημειώνεται ως ***an(α)*** ,όταν υπάρχει ένα προσανατολισμένο μονοπάτι (***directed path***) από το ***υ*** στο ***α***.Αντίστροφα,υπάρχει ένα προσανατολισμένο μονοπάτι από την *α* σε κάθε από τους απογονούς της (***descendants***) Στο σχήμα 2.2,για παράδειγμα η μεταβλητή ***d*** έχει γονείς τις μεταβλητές ***a ,b*** δηλαδή $pa_m(a)=\{a,b\}$,και παιδιά της μεταβλητές ***f ,g.*** Στους γείτονες της ***d*** περιλαμβάνονται οι ***a,b,f*** και ***g*** στον σκελετό $\bar{m}$ Επιπλέον ,η μεταβλητή ***a*** έχει ως απόγονους τις ***c,d,f*** και την ***g***,ενώ οι πρόγονοι της ***g*** είναι το σύνολο $an(g) = \{a,b,d,e\}$

Λόγω της ακυκλότητας(***acyclicity***) του, ένα ***DAG*** συνεπάγεται μια *προγονική εντολή* ***(ancestral ordering****) πάνω στις μεταβλητές.*



Σχήμα 2.2 :Ένα απλοϊκό ***DAG*** ,ο σκελετός του $\tilde{m}$ και ένα σχέδιο ισοδύναμης κατηγορίας

Αυτή είναι μια *συνολίκη σχέση εντολής* **(*ordering relation*** $\prec$ **)** ,δηλαδή για όλες τις $a,b, \in V$ πρέπει να ισχύει ότι: (1) $a \prec b \; V \; b \prec a$, (2) $a \prec b \wedge b \prec a$ τότε συνεπάγεται ότι $a=b$(3) $a \not\prec a$, and (4) $a \prec b \wedge b \prec c$ implies $a \prec c$.Τυπικά ,ένα ***DAG*** δεν ορίζει μια μοναδική ολική διαταγή . Απλά μόνο ορίζει μια μερική(partial).Στο σχήμα 2.2 δύο έγκυρες προγονκές διαταγές μεταξύ των άλλων είναι:. $a \prec b \prec e \prec d \prec e \prec f \prec g$ ή $b \prec a \prec d \prec f \prec e \prec g \prec c$.

Υπάρχουν ουσιαστικά δύο εναλλακτικές λύσεις για το πώς να συσχετίσθουν η γραφική δομή ενός Μπεϋζιανού δικτύου με τις υπό όρους ανεξαρτησίες και τις εξαρτήσεις που κρύβονται κάτω από τις κατανομές πιθανότητας, οι ***markov assumptions*** και το ***d-separation criterion***.Όσον αφορά την πρώτη , τρεις παραλλαγές πρέπει να διακριθούν: ***directed pairwise***, ***directed local*** και τις ***directed global Markov properties*** Εάν η κατανομή πιθανότητας είναι αυστηρά θετική, και οι τρεις markov ιδιότητες μπορούν να αποδειχθούν ότι είναι ισοδύναμες. Η markov assumption εφαρμόζεται σε πολλούς τομείς της έρευνας προκειμένου να προσεγγιστούν τα προβλήματα που είναι αρκετά πολύπλοκα.

Η markov προσέγγιση λέει ιδανικά ότι η κατάσταση μιας μεταβλητής εξαρτάται μόνο από την κατάσταση που λαμβάνεται από τις μεταβλητές στην εγγύτητά της, δηλ. τα τελευταία προστατεύουν αυτήν την μεταβλητή από την επιρροή των άλλων μεταβλητών στην περιοχή. Παραδείγματος χάριν, στην ανάλυση χρονοσειρών (***time-series***) συχνά υποτίθεται ότι μόνο η τρέχουσα κατάσταση του κόσμου ασκεί επίδραση στην κατάσταση του κόσμου στο επόμενο χρόνος-βήμα, ανεξάρτητα από το παρελθόν.

Τώρα εστιάσουμε στο δεύτερο κριτήριο, δηλ. το ***d-separation criterion***.Μποεί να αποδειχθεί ότι το ***d-separation criterion*** είναι ισοδύναμο με την ***directed global Markov*** ιδιότητα των Μπαεζιανών δικτύων, και ως εκ τούτουκαι με τις άλλες markov ιδιότητες υπό τον όρο ότι η κατανομή πιθανότητας είναι αυστηρά θετική.

Ορισμός 2.1 (D-Separation) Μέσα στο ***DAG*** δύο ασυσχέτιστα (disjoint) σύνολα ***A*** και ***B*** χωρίζονται από ένα τρίτο σύνολο $S \subseteq V \backslash (AUB)$ *,που συμβολίζεται με* $A \perp B \mid S$, *αν και μόνο αν κατά μήκος κάθε μονοπατιού μεταξύ των* ***A*** *και* ***B*** *υπάρχει μια μεταβλητή s η οποία ικανοποιεί μία από από τις δύο συνθήκες:*

- το ***s*** έχει συγκλινόμενες ακμές και κανένα από τα s ή των απγόνων του δεν ανήκουν στο ***S***

- το s δεν έχει συγκλίνοντες ακμές και το $s \in S$.

Μια μεταβλητή $s \in V$ έχει συγκλίνοντες ακμές $"\rightarrow s \leftarrow"$, όταν η προηγούμενη (preceding)και διαδοχική(successive) μεταβλητή κατά μήκος του μονοπατιού είναι και οι δύο γονείς της spreceding and the successive variable along the path are both parents of s Όχι μόνο η παρουσία ακρών αλλά και οι προσανατολισμοί τους είναι κρίσιμοι προκειμένου να δώσει τις μεταβλητές ***d-separated***. Αυτός ο καθορισμός υπονοεί ότι μια πορεία μεταξύ δύο μεταβλητών *εμποδίζεται* όταν τηρείται ένας από τους ανωτέρω δύο όρους, και *ενεργοποιείται* ειδάλλως. Ως εκ τούτου, δύο μεταβλητές $a, b \in V$ είναι ***d-separated*** από το S, όταν εμποδίζονται όλες οι πορείες μεταξύ τους. Είναι σημαντικό να σημειώσουμε ότι $A \perp B \mid S$, δεν σημαίνει ότι $a \perp b \mid S'$ για $S' \neq S$.Αύτο ισχύει ακόμα και όταν $S' \supset S$ όταν ένα μπλοκαρισμένο μονοπάτι μπορεί μετατραπεί σε ενεργό από μια επιπρόσθετη μεταβλητή since $s' \in S'$, $s' \notin S$ μέσω της πρώτης συνθήκης που αφέρθηκε προηγουμένως. Αυτό είναι μια σημαντική διαφορά με τα Μαρκοβιανά μοντέλα.Στο ***DAG*** που παρουσιάζεται στο σχήμα 2.2 για παράδειγμα οι μεταβλητές ***d*** και ***e*** είναι ***d-separated*** δεδομένου ***του b***,αλλά δεν είναι ***d-separated*** με γνωστά τα ***b*** και ***g***.Αν δύο ασύνδετα σύνολα ***A*** και ***B*** δεν είναι ***d-separated*** από ένα σύνολο ***S*** τότε καλούνται ***d-connected*** και συμβολίζεται ως $A \not\perp B \mid S$.

Τώρα θα συνδέσουμε τη δομή ενός Bayesian network με το σύνολο των κατανομών πιθανότητας που περιγράφεται με τη βοήθεια του d-separation criterion. Ειδικότερα ,μια **d-separation** $a \perp b \mid S$ αναγνωρίζουν ότι ένα ***DAGm*** περιέχει την κατά συνθήκη ανεξαρτησία $a \perp b \mid S$ μες στην κατανομή πιθανότητας που περιγράφηκε προηγουμένως. Με άλλα λόγια, όλες οι κατανομές πιθανότητας εκθέτουν τις υπό όρους ανεξαρτησίες που υπονοούνται από το ***DAG m***, αλλά διαφέρουν η μια από την άλλη λόγω των διαφορετικών τιμών των παραμέτρων θ που επιλέγονται στο Μπαεζιανό δίκτυο.Οι συγκεκριμένες επιλογές των παραμέτρων μπορούν να εμπεριέχουν τις κατανομές πιθανότητας που υπονοούν τις πρόσθετες ανεξαρτησίες που δεν αντιπροσωπεύονται στο ***DAG.***Εντούτοις, μπορεί να αποδειχθεί ότι *σχεδόν όλες* οι κατανομές πιθανότητας που περιγράφονται από τα Μπαεζιανά δίκτυα (υπό μια μέτρο-θεωρητική έννοια-measure-theoretic sense) υπονοούν μια υπό όρους ανεξαρτησία εάν και μόνο εάν το ***DAG*** αντιπροσωπεύουν τον αντίστοιχο ***d-separated.***

2.4.2Επαναλαμβανόμενη παραγοντοποίηση της κατανομής πιθανότητας (Recursive Factorization of the Probability Distribution)

Το κριτήριο ***d-separated*** και οι markov ιδιότητες είναι ισοδύναμες με ακόμα ένα χαρακτηριστικό των Μπεϋζιανών δικτύων, δηλαδή στη αποσυνθετικότητα (decomposability) της κατανομής πιθανότητάς του όταν το τελευταίο είναι αυστηρά θετικό. Ένα Μπεϋζιανό μοντέλο δικτύων με το ***DAG m*** περιγράφει μια κατανομή πιθανότητας για ένα σύνολο μεταβλητών V που παραγοντοποιεί κατ' επανάληψη όπως:

$$p(V) = \prod_{v \in V} p(v \mid pa_m(v)) \qquad (2.2.1)$$

Η πολλαπλο-μεταβλητή κατανομή πιθανότητας(***multivariate probability distribution***)για το σύνολο $\boldsymbol{\nu}$ εκ ως εκ τούτου αποσυνθέτει σε ***univariate*** κατανομές πιθανότητας, δίνοντας ένα μοντέλο Bayesian δικτύου το οποίο είναι τροπικό(modular).Το σύνολο των παραμέτρων $\boldsymbol{\theta}$ ενός μοντέλου Μπαεζιανού δικτύου είναι ένα σύνολο των κατά συνθήκη πιθανοτήτων $p(v \mid \mathrm{pa}_m(v))$ όπου $\mathrm{pa}_m(v)$ δηλώνει τους γονείς της μεταβλητής υ στο ***DAG m.***Κατ' αρχάς, δεδομένου ότι κάθε μια από τις υπό όρους πιθανότητες περιλαμβάνει χαρακτηριστικά μόνο έναν μικρό αριθμό μεταβλητών, δηλ. $|\mathrm{pa}_m(v)| \ll |\mathbf{V}|$ για κάθε $v \in \mathbf{V}$, οι παράμετροι ενός Μπαεζιανού δικτύου μπορούν να υπολογιστούν από τα πεπερασμένα στοιχεία με την αυξανόμενη αξιοπιστία. Δεύτερον, οι παράμετροι ενός Μπαεζιανού δικτύου, δεδομένου ότι είναι υπό όρους πιθανότητες, μπορούν άμεσα να υπολογιστούν από τη κατανομή πιθανότητας που υπονοείται από τα στοιχεία. Αντίθετα, η εκτίμηση παραμέτρου σε άλλα γραφικά μοντέλα απαιτεί γενικά τις επαναληπτικές διαδικασίες, π.χ.το επαναληπτικό αναλογικό ξελέπιασμα(scaling) των Markov δικτύων

2.4.3 Μαρκοβιανή ισοδυναμία(Markov Equivalence)

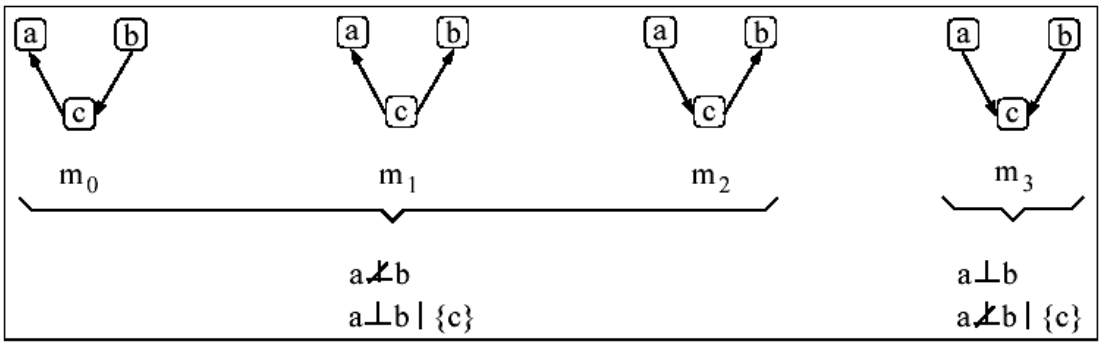
Ένα ***DAG*** καθορίζει ένα μοναδικό σύνολο κατανομών πιθανότητας που εκφράζουν τις υπό όρους ανεξαρτησίες σύμφωνα με το ***d-separation*** criterion . Η αντίθετη κατάσταση αντιμετωπίζεται στη δομική εκμάθηση, όπου η ερώτηση προκύπτει εάν οι υπό όρους ανεξαρτησίες και οι εξαρτήσεις που υπονοούνται από μια κατανομή πιθανότητας καθορίζουν ένα μοναδικό ***DAG***. Γενικά, η απάντηση είναι αρνητική. Παραδείγματος χάριν, τα ***DAGs*** $\boldsymbol{m_0}$, $\boldsymbol{m_1}$ και $\boldsymbol{m_2}$ στο σχήμα 2.3 επιδεικνύουν τους ίδιους ***d-separations*** και τις ***d-connections***. Ως εκ τούτου, περιγράφουν το ίδιο σύνολο κατανομών πιθανότητας. Αυτό μπορεί επίσης να φανεί από την παραγοντοποίηση της κατανομής πιθανότητας (βλ. εξίσωση 2.2.1), δεδομένου ότι είναι ισοδύναμη με το ***d-separations*** criterion δεδομένων αυστηρά των θετικών κατανομών. Οι ακόλουθες τρεις παραγοντοποιήσεις συνεπάγονται από το ***DAGs*** $\boldsymbol{m_0}$, $\boldsymbol{m_1}$ και $\boldsymbol{m_2}$ στο σχήμα 2.3:

$$p(a,b,c) = \underbrace{p(a|c)\,p(c|b)\,p(b)}_{m_0} = \underbrace{p(a|c)\,p(b|c)\,p(c)}_{m_1} = \underbrace{p(c|a)\,p(b|c)\,p(a)}_{m_2} \qquad (2.2.2)$$

Η ταυτότητα των διαφορετικών παραγοντοποιήσεων είναι προφανής, μια και το δεύτερο προϊόν προκύπτει από την πρώτη λόγω του γενικού νόμου:

$$p(c|b)\,p(b) = p(b,c) = p(b|c)\,p(c).$$

και ομοίως ο τρίτος από τον δεύτερο, υποθέτοντας $\boldsymbol{p(.)>0}$. Αν και το ***DAG*** $\boldsymbol{m_3}$ περιέχει τις ίδιες ακμές με τα προηγούμενα τρία ***DAGs,*** αντιπροσωπεύει τους διαφορετικούς ***d-separations*** και ***d-connections*** . Αυτό είναι επειδή το ***DAG*** $\boldsymbol{m_3}$ έχει ένα ***collider*** στο μεταβλητή ***c***. Ένα ***collide***r, ή η ***v-δομή***, είναι ένα διαταγμένη τριάδα των μεταβλητών $\boldsymbol{a}$, $\boldsymbol{c}$, $\boldsymbol{b} \in \boldsymbol{\nu}$ έτσι ώστε η ακμή μεταξύ του ***a*** και του ***b*** είναι απούσα και $a \rightarrow c \leftarrow b$



Σχήμα 2.3:Τα ***DAGs*** m_0, m_1, m_2 ανήκουν σε ισοδύναμες κατηγορίες σε αντίθεση με το m_3 που ανήκει σε διαφορετική

Είναι προφανές ότι οι ***d-separations*** και οι ***d-connections*** καθιερώνουν μια αντανακλαστική, συμμετρική και μεταβατική σχέση μεταξύ ***DAGs***, το οποίο καλείται ***markov equivalence relation*** Αυτό επιτρέπει να χωρίσει το διάστημα ***DAGs*** *σε ισοδύναμες κατηγορίες*, όπου κάθε κατηγορία ισοδυναμίας περιέχει όλα τα ***DAGs*** που είναι markov ισοδύναμα το ένα στο άλλο. Μπορεί να αποδειχθεί ότι δύο ***DAGs*** είναι ισοδύναμα εάν και μόνο εάν αυτοί έχουν τον ίδιο σκελετό και τα ίδια ***colliders***. Αυτό σημαίνει ότι όλες οι ακμές που δεν περιλαμβάνονται σε ένα ***collide***r μπορούν να προσανατολιστούν αυθαίρετα εφ' όσον δεν εμφανίζεται ένα πρόσθετο ***collider***. Οι άκρες μπορούν ως εκ τούτου να διαιρεθούν σε αυτές με αναστρέψιμο προσανατολισμό και σε άλλες με αμετάκλητο. Οι άκρες σ' έναν αμετάκλητο προσανατολισμό καλούνται επίσης *αναγκασμένες* ακμές(***compelled***). Μια κατηγορία ισοδυναμίας απεικονίζεται συχνά με τη βοήθεια ενός αποκαλούμενου "*σχέδιο"*, όπου οι ακμές με έναν αναστρέψιμο προσανατολισμό επιδεικνύονται χωρίς κατευθύνσεις έτσι ώστε μόνο οι αμετάκλητοι προσανατολισμοί παρουσιάζονται. Αυτό είναι διευκρινισμένο στο σχήμα 2.2.

Η ύπαρξη των κατηγοριών ισοδυναμίας έχει τρεις συνέπειες σχετικά με τη δομική εκμάθηση/εκπαίδευση των Μπαεζιανών δικτύων. Πρώτος και ο σημαντικότερος λόγος,είναι ότι μόνο η κατηγορία ισοδυναμίας παρά ένα ιδιαίτερο ***DAG*** μπορεί γενικά να προκληθεί/δημιουργηθεί από τη κατανομή πιθανότητας που υπονοείται από τα δεδομένα. Μόνο εάν οι στατιστικοί αλγόριθμοι εκμάθησης συμπληρώνονται με την επιπρόσθετη γνώση, π.χ. για τους προσανατολισμούς των ακμών, ένα ιδιαίτερο ***DAG*** μπορεί να καθοριστεί. Δεύτερον, οι *undirected/απροσανοτολισμένες* ακμές στο προκληθέν σχέδιο της κατηγορίας ισοδυναμίας δεν μπορούν σαφώς να ερμηνευθούν ως ένωση μεταξύ μιας αιτίας ν και μιας συνέπειας εάν καμία πρόσθετη γνώση δεν είναι διαθέσιμη που υπονοεί ορισμένους προσανατολισμούς. Τρίτον, η δομική εκμάθηση στο διάστημα αναζήτησης ***DAGs*** πάσχει από μερικά μειονεκτήματα, τα οποία μπορούν να υπερνικηθούν στο διάστημα αναζήτησης των κατηγοριών ισοδυναμίας(***equivalence classes***). Οι αλγόριθμοι εκμάθησης που λειτουργούν στο διάστημα των κατηγοριών ισοδυναμίας είναι, εντούτοις, υπολογιστικά πολύπλοκοι .Η προσέγγιση βασισμένη σε περιορισμούς είναι ένας αποδοτικός τρόπος να προκληθεί/αποδοθεί άμεσα το σχέδιο(pattern) μιας κατηγορίας ισοδυναμίας

Εάν κάποιος επιθυμεί να εξερευνήσει διάφορα ***DAGs*** που περιλαμβάνονται στην ίδια κατηγορία ισοδυναμίας, η έννοια μιας "*καλυμμένης" (covered)*ακμής είναι χρήσιμη. Μια προσανατολισμένη άκρη μεταξύ δύο μεταβλητών $a, b \in \mathbf{V}$ λέγεται ότι

είναι καλυμμένο σε ένα ***DAGm*** εάν υποστηρίζει ότι: $\mathbf{pa}_m(a) \setminus \{b\} = \mathbf{pa}_m(b) \setminus \{a\}$, δηλ. κατά την αμέλεια της ακμής μεταξύ a και b έχουν τους ίδιους γονείς. Ο προσανατολισμός μιας καλυμμένης άκρης μπορεί να αναστραφεί προκειμένου να ληφθεί ένα άλλο ισοδύναμο ***DAG.*** Αυτό έχει τις σημαντικές συνέπειες για τις ιδιότητες της συνάρτησης αποτελέσματος.

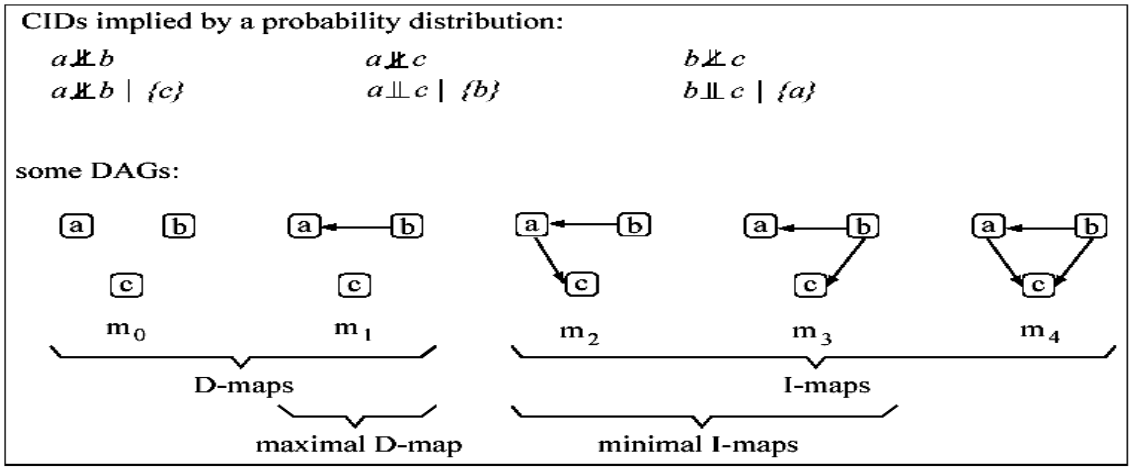
2.4.4 Perfect Map and Faithfulness

Η δομική εκμάθηση στοχεύει στην πρόκληση ***DAGs*** που περιγράφουν μια δεδομένη κατανομή πιθανότητας που υπονοείται από τα δεδομένα.. Αυτό το τμήμα εξετάζει έτσι την ερώτηση εάν μια δομή Μπαεζιανών δικτύων είναι σε θέση να περιγράψει όλες τις υπο όρους ανεξαρτησίες και εξαρτήσεις(***CIDs)*** που υπονοούνται από ένα αυθαίρετο σύνολο δεδομένων.Παραδείγματος χάριν,υποθέστε ότι μια κατανομή πιθανότητας για τις τρεις μεταβλητές ***a,b***και το ***c*** υποδηλώνει τις υπό όρους ανεξαρτησίες $a \perp\!\!\!\perp c \mid \{b\}$ and $b \perp\!\!\!\perp c \mid \{a\}$. ενώ οι άλλες ενώσεις είναι εξαρτήσεις.Προφανώς το ***CIDs*** δεν μπορεί να αντιπροσωπευθεί σε ένα ενιαίο ***DAG*** (βλέπε σχήμα 2.4).Αυτό προτείνει να αντιπροσωπεύσει είτε όλες τις (υπο όρους) εξαρτήσεις σε ένα ***DAG***.Αυτό οδηγεί στην έννοια των I-maps .

Ορισμός 2.2 (Independence Map (I-Map))

Ένα ***DAG*** *είναι ένας ανεξάρτητος χάρτης(**independence map**) μιας κατανομής πιθανότητας για ένα σύνολο μεταβλητών* $\boldsymbol{\mathcal{V}}$ να και μόνο εάν : $a \perp b \mid \mathcal{S} \Rightarrow a \perp b \mid \mathcal{S}$, *ή ισοδύναμα:* $a \not\perp b \mid \mathcal{S} \Rightarrow a \not\perp b \mid \mathcal{S}$ (για όλα $a, b \in \mathcal{V}, \mathcal{S} \subseteq \mathcal{V} \setminus \{a, b\}$)

Αυτό σημαίνει ότι κάθε ***d-separation*** που επιδεικνύεται σε έναν ***I-map*** συνεπάγεται μια (υπό όρους) ανεξαρτησία στη κατανομή πιθανότητας. Εντούτοις, η κατανομή πιθανότητας να υπονοήσει τις πρόσθετες ανεξαρτησίες που δεν αντιπροσωπεύθηκαν σε έναν ***I-map***. Προφανώς, η πλήρης γραφική παράσταση, δηλ. αυτή όπου όλες οι άκρες είναι παρούσες, είναι ένας τετριμμένος ***I-map*** οποιασδήποτε κατανομής πιθανότητας. Πιό ενδιαφέροντες είναι *οι* ***minimal I-maps***, οι οποίοι είναι ελάχιστοι υπό την έννοια ότι καμία άκρη δεν μπορεί να αφαιρεθεί χωρίς καταστροφή της ιδιότητα , της ύπαρξης ενός ***I-map***. Γενικά, μπορεί να υπάρξουν *διάφοροι* ***minimal I-maps***, που δεν είναι markov ισοδύναμο.. Αυτό είναι διευκρινισμένο στο σχήμα 2.4, απεικονίζοντας δύο ***minimal I-maps*** που περιέχουν τις διαφορετικές ακμές.



Σχήμα *2.4:Ένα* ***DAG*** *μπορεί να μην είναι σε θέση να αναπαραστίσει όλες τις* ***CIDs*** *που δημιουργούνται από μια αυθαίρετη κατανομή πιθανότητας*

Ορισμός 2.3 (Dependence Map (D-map)) *Ένα* ***DAG*** *είναι ένας ΄΄* ***dependence map΄΄*** *μιας κατανομής πιθανότητας για ένα σύνολο μεταβλητών* $\boldsymbol{\mathcal{V}}$ *αν και μόνο εάν ισχύει:* $a \not\perp b|\mathcal{S} \Rightarrow a \not\perp b|\mathcal{S}$ *ή* $a \perp b|\mathcal{S} \Rightarrow a \perp b|\mathcal{S}$,(για όλα $a, b \in \mathcal{V}, \mathcal{S} \subseteq \mathcal{V} \setminus \{a, b\}$)

Ένας ***D-map*** ως εκ τούτου αντιπροσωπεύει όλες τις υπό όρους ανεξαρτησίες που κρύβονται κάτω από τη κατανομή πιθανότητας, αλλά οι πρόσθετες ανεξαρτησίες μπορούν ενδεχομένως να "διαβαστούν" από έναν ***D-map***. Ισοδύναμα, όλες οι εξαρτήσεις που υπονοούνται από έναν ***D-map*** είναι παρούσες στη κατανομή πιθανότητας. Κατά συνέπεια, η κενή γραφική παράσταση είναι ένας τετριμμένος(trivial) ***D-map***, δεδομένου ότι δεν αντιπροσωπεύει οποιεσδήποτε εξαρτήσεις. Ένα ***DAG*** είναι ένας ***maximal D-map*** εάν καμία ακμή δεν μπορεί να περιληφθεί στη γραφική παράσταση χωρίς απώλεια της ιδιότητας της ύπαρξης ενός ***D-map***. Όπως στους ***I-maps*** ,μπορούν γενικά να υπάρξουν διάφοροι ***D-maps*** που δεν είναι ισοδύναμοι ο ένας με τον άλλο. Δεδομένου ότι οι ***D-maps*** αντιπροσωπεύουν λιγότερες εξαρτήσεις, περιέχουν χαρακτηριστικά έναν μικρότερο αριθμό ακμών σε σχέση από τους ***I-maps***. Στην ειδική περίπτωση ότι μια κατανομή πιθανότητας είναι τέτοια που ένα ***DAG*** μπορεί ταυτόχρονα να είναι ένας ελάχιστος ***minimal I-map***και ένας maximal ***D-map*** της κατανομής, αυτό το ***DAG*** καλείται **perfect map**, και η κατανομή πιθανότητας καλείται πιστή.(faithful)

Ορισμός 2.4 (Perfect Map and Faithfulness)Ένα ***DAG*** *είναι ένας* ***perfect map(τέλειος χάρτης)*** *της κατανομής πιθανότητας για ένα σύνολο μεταβλητών* $\boldsymbol{\mathcal{V}}$ *αν και μόν εάν* $a \perp\!\!\!\perp b|\mathcal{S} \Leftrightarrow a \perp b|\mathcal{S}$ *(για όλες τις* $\boldsymbol{a, b \in \mathcal{V}}$ $\mathcal{S} \subseteq \mathcal{V} \setminus \{a, b\}$*)Μια κατανομή πιθανότητας είναι ΄΄πιστή΄΄ αν και μόνο εάν έχει ένα* ***perfect map****.*

Εάν μια κατανομή πιθανότητας είναι "πιστή" τότε ο ***perfect map*** καθορίζεται *μεμονωμένα* (με Markov ισοδυναμία, φυσικά), αντίθετα από τους ***I-maps*** και τους ***D-maps*** . Η προσέγγιση βασισμένη σε περιορισμούς στη δομική εκμάθηση στα Μπαεζιανά δίκτυα απαιτεί χαρακτηριστικά την υπόθεση ότι η κατανομή πιθανότητας που υπονοείται από τα δεδομένα είναι πιστή.

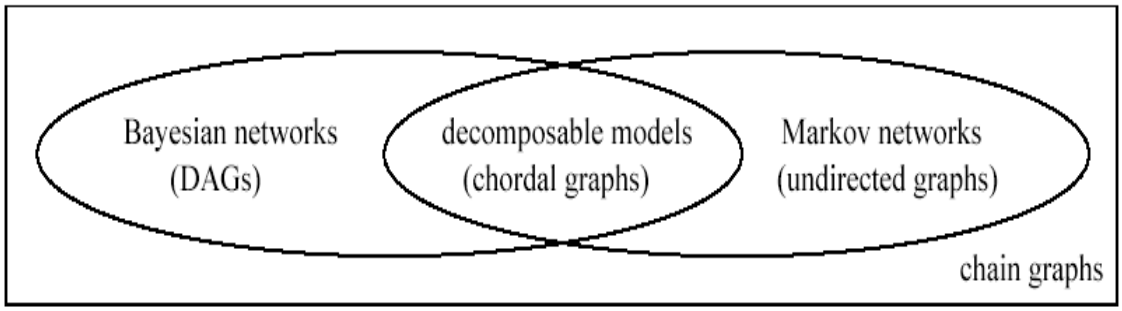
2.5 Graphical Models

Τα γραφικά μοντέλα μπορούν να πρωτοεμφανίστηκαν κάπου στην αρχή του εικοστού αιώνα, όπου χρησιμοποιήθηκαν αρχικά στη στατιστική φυσική. Ανεξάρτητα, έχουν εξελιχθεί στη γενετική και τις στατιστικές. Στις στατιστικές, τα αποκαλούμενα *λογαριθμικά-γραμμικά μοντέλα(****log-linear models****)* είναι από καιρό μια δημοφιλής προσέγγιση στη διαμόρφωση των πολλών μεταβλητών κατανομών πιθανότητας για τις διακριτές μεταβλητές. *Τα γραφικά μοντέλα* μπορούν να θεωρηθούν ως ειδική περίπτωση των αποκαλούμενων *ιεραρχικών* λογαριθμικό-γραμμικών μοντέλων(*hierarchical* log-linear models.Η.κατηγορία των γραφικών μοντέλων περιλαμβάνει όχι μόνο τα Μπαεζιανά δίκτυα αλλά και τα markov δίκτυα και τις γραφικές παραστάσεις αλυσίδων(***chain graphs***)Ενώ τα τελευταία δύο μοντέλα χρησιμοποιούνται ευρέως στις στατιστικές, τα Μπαεζιανά δίκτυα έχουν προσελκύσει πολλή προσοχή στην κοινότητα AI κατά τη διάρκεια των προηγούμενης μιας ή δύο

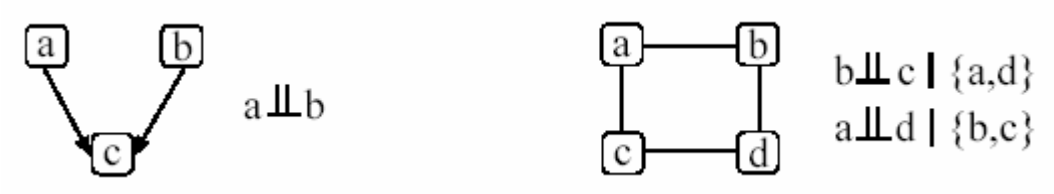
δεκαετιών. Σε αντίθεση με τα Μπαεζιανά δίκτυα, τα markov δίκτυα περιέχουν απλώς απροσανατολισμένες ακμές (***undirected edges***). Δεδομένου ότι οι γραφικές παραστάσεις αλυσίδων επιτρέπουν και τις προσανατολισμένες όσο και τις undirected ακμές, τα Μπαεζιανά δίκτυα και τα markov δίκτυα μπορούν να αντιμετωπισθούν ως ειδικές περιπτώσεις των γραφικών παραστάσεων αλυσίδων. Το κοινό χαρακτηριστικό των διάφορων γραφικών μοντέλων είναι η γραφική απεικόνιση των υπό όρους ανεξαρτησιών και των εξαρτήσεων (***CIDs***) με τη βοήθεια των markov ιδιοτήτων. Όσον αφορά ένα markov δίκτυο, μια υπό όρους ανεξαρτησία $\alpha \perp\!\!\!\perp b \mid S$ αντιστοιχεί σε έναν χωρισμό της ***α*** και της ***b*** στη undirected γραφικής παράστασης έτσι ώστε κάθε πορεία μεταξύ της ***α*** και του ***b*** περιέχει μια μεταβλητή στο ***S.*** Αυτό συνεπάγεται πάλι μια παραγοντοποίηση της κοινής κατανομής πιθανότητας, παρόμοια με τα Μπαεζιανά δίκτυα. Εντούτοις, δεν μπορεί να εκφραστεί από την άποψη (των υπό όρους) πιθανοτήτων γενικά. Αυτό δίνει την εκτίμηση της παραμέτρου στα markov δίκτυα αρκετά πολύπλοκα, δεδομένου ότι εκείνοι οι παράγοντες της κατανομής πιθανότητας, οι αποκαλούμενες clique δυνατότητες, δεν μπορούν να υπολογιστούν άμεσα από τα στοιχεία, αλλά συνήθως απαιτεί μερικά επαναληπτικά σχέδια όπως το επαναληπτικό ανάλογο ξελέπιασμα(scaling).

Το σχήμα 2.5 απεικονίζει τις κατηγορίες κατανομών πιθανότητας με τους όρους των ***CIDs*** που μπορεί να συλληφθεί από τα Μπαεζιανά δίκτυα και τα markov δίκτυαΕίναι προφανές ότι ορισμένες ***CIDs*** μπορεί μόνο να αντιπροσωπευθούν από ***DAGs*** ενώ άλλες (εξαρτήσεις και ανεξαρτησίες) μπορούν μόνο να απεικονιστούν από τη undirected γραφική παράσταση ενός markov δικτύου. Ένα παράδειγμα για κάθε ένα του οποίου απεικονίζεται στο σχήμα 2.6. Επιπλέον, υπάρχει επίσης μια επικάλυψη και μεταξύ των δύο μοντέλων, που δείχνουν ότι ορισμένες κατανομές πιθανότητας μπορούν να περιγραφούν και από ένα Μπαεζιανό δίκτυο όπως επίσης και από ένα markov δίκτυο. Πολλοί δομικοί αλγόριθμοι εκμάθησης (structural learning algorithms)εστιάζουν στη δημιουργία αυτών των αποκαλούμενων *μοντέλων αποσύνθεσης(**decomposable models)***, δεδομένου ότι έχουν μερικές ιδιότητες που διευκολύνουν την επαγωγή τους από τα στοιχεία. Η δομή των decomposable μοντέλων απεικονίζεται χαρακτηριστικά από μια (undirected) χορδική (***chordal***) γραφική παράσταση. Μια undirected γραφική παράσταση καλείται ***χορδική***, ή ***triangulated***, εάν κάθε κλειστή πορεία (βρόχος) του μήκους τεσσάρων ή περισσοτέρων έχει τουλάχιστον μια χορδή, δηλ. μια ακμή μεταξύ δύο μεταβλητών κατά μήκος του βρόχου που δεν περιλαμβάνεται στο βρόχο.

Αυτό παράγει ότι η undirected γραφική παράσταση στο σχήμα 2.6 δεν είναι **chordal.** Ισοδύναμα σε μια (*undirected) chordal* γραφική παράσταση, η δομή μπορεί επίσης να αντιπροσωπευθεί από ένα DAG χωρίς ***colliders***.Αυτό δείχνει ότι η κατανομή πιθανότητας που περιγράφεται από ένα ***decomposable*** μοντέλο παραγοντοποιεί τις πιθανότητες ενός δημιουργήματος των υπό συνθήκη πιθανοτήτων που είναι παρόμοιο με τα Μπαεζιανά δίκτυα, ενώ η γραφική παράστασή της μπορεί παρουσιαστεί με τη χρήση των undirected ακμών , που αντιστοιχούν σε ένα markov δίκτυο.



Σχήμα 2.5: Ομοιότητα των γραφικών μοντέλων



Σχήμα 2.6*:*Το ***CIDs*** που αντιπροσωπεύεται από το ***DAG*** (αριστερά) δεν μπορεί ολόκληρο να επιδειχθεί στη undirected γραφική παράσταση ενός markov δικτύου. Αντιθέτως, δεν υπάρχει κανένα ***DAG*** που μπορεί να αντιπροσωπεύσει όλο το ***CIDs*** που παρουσιάζεται από τη markov δομή δικτύων (δεξιά).

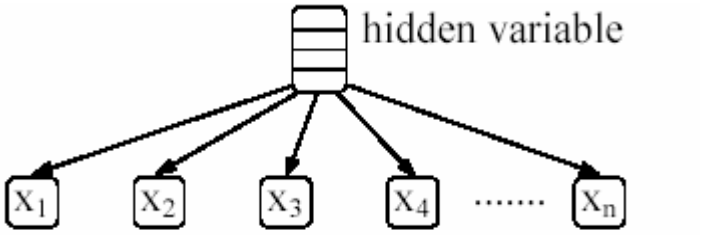
Τέλος,ας σημειώσουμε ότι οι κατανομές πιθανότητας που περιγράφονται από τα γραφικά μοντέλα ανήκουν στην εκθετική οικογένεια. Λεπτομερώς, τα γραφικά μοντέλα με τις κρυμμένες μεταβλητές, δηλ. μεταβλητές που είναι απαρατήρητες, ανήκουν στην αποκαλούμενη στρωματοποιημένη εκθετική οικογένεια.(***stratified exponential family***). Ελλείψει των κρυμμένων μεταβλητών, τα Μπαεζιανά δίκτυα είναι κυρτές εκθετικές οικογένειες, ενώ τα markov δίκτυα είναι γραμμικές εκθετικές οικογένειες.

2.6 Μερικά σχετικά εργαλεία για την ανάλυση

Υπάρχουν διάφορες προσεγγίσεις στην ανάλυση δεδομένων.Ας . απεικονίσουμε δύο από τους όποιους συσχετίζονται πολύ με τα Μπαεζιανά δίκτυα. Οι κανόνες ένωσης(association rules) είναι μια δημοφιλής τεχνική στην ανάσυρση δεδομένων(***data mining***). Χρησιμοποιούνται στον υπολογισμό της δύναμης των στατιστικών συσχετίσεων μεταξύ των μεταβλητών.

Στην περιοχή που απεικονίζεται στο σχήμα 2.1, οι μεταβλητές " αποτυχία φωτισμού;" και "αποτυχία του υπολογιστή;"είναι μόνο έμμεσα συσχετισμένες. Δεδομένου ότι οι κανόνες ένωσης δεν μπορούν να αποτελέσουν τέτοιες *υπό όρους* ανεξαρτησίες, παράγουν, μια ενδεχομένως ισχυρή, στατιστική ένωση μεταξύ της " αποτυχίας του φωτισμού;" και της "αποτυχίας του υπολογιστή;". Κάποιος μπορεί να φανταστεί ότι, στις περιοχές /***domains*** με πολλές μεταβλητές, μερικές άμεσες ενώσεις συνεπάγονται έναν μεγάλο αριθμό έμμεσων ενώσεων. Προκειμένου να αποκτηθεί γνωση σε μια τέτοια περιοχή είναι έτσι επιθυμητό να διακρίνει μεταξύ των άμεσων

και έμμεσων ενώσεων, δεδομένου ότι τα πρώτα συνεπάγονται τα τελευταία. Για αυτόν τον λόγο, οι κανόνες ένωσης παρέχουν τη λιγότερη γνώση στις αμοιβαίες σχέσεις των διάφορων μεταβλητών σε μια περιοχή απ΄ότι τα Μπεϋζιανά δίκτυα, ή τα γραφικά μοντέλα, γενικά.



Σχήμα 2.7: Ένα naive Bayesian network μπορεί να χρησιμοποιηθεί για clustering..

Επίσης θα αναφερθούμε στηκατηγοριοποίηση (***clustering***), άλλη μια δημοφιλής προσέγγιση στην ανάλυση δεδομένων, η οποία συσχετίζεται στενά με ένα ιδιαίτερο είδος Μπαεζιανών δικτύων, το αποκαλούμενο αφελές Μπαεζιανό δίκτυο (**naive Bayesian network**) Μέσω του **clustering** στοχεύεται η ομαδοποίηση των παρόμοιων στοιχείων που περιλαμβάνονται στη βάση δεδομένων στην ίδια συστάδα(cluster). Τα δεδομένα μπορούν να θεωρηθούν ως ένα σύνολο περιπτώσεων, κάθε ένα από το οποίο που αντιπροσωπεύει μια διαμόρφωση(configuration). Η ομοιότητα των διαμορφώσεων μπορεί να μετρηθεί από έναν μετρht;h (metric)που καθορίζεται στο διάστημα των διαμορφώσεων. Το τελευταίο καλείται ***feature space in the context of clustering.*** Το διάστημα χαρακτηριστικών γνωρισμάτων ως εκ τούτου χωρίζεται ανεξάρτητα υποδιαστήματα (***subspaces***), όπου κάθε ***subspace*** καλείται μια συστάδα ή σενάριο(***scenario***).

Χαρακτηριστικά, κάθε διαμόρφωση ανήκει σε μια ορισμένη cluster , και το σύνολο των δεδομένων μπορεί ως εκ τούτου να γίνει κατανοητό από την σκοπιά (μερικών) σεναρίων, συλλαμβάνοντας τα κύρια αποτελέσματα που κρύβονται κάτω από τα στοιχεία. Εάν το σύνολο των δεδομένων είναι τέτοιο που τα σενάρια δεν είναι καλά-χωρισμένα το ένα από το άλλο, και κάποιο προτιμήσει *τη «μαλακή»* συγκέντρωση(***soft clustering***), όπου μια περίπτωση είναι ελαχίστως (fractionally)ορισμένη στις πολλαπλάσιες clusters Εκτός από άλλες μεθόδους, αυτό μπορεί να επιτευχθεί με τη βοήθεια ενός ***naive Bayesian network***.Περιέχει μια κρυμμένη, διακριτή μεταβλητή της οποίας οι καταστάσεις αντιστοιχούν στις διαφορετικά clusters ή τα σενάρια. Επιπλέον, οι διαφορετικές μεταβλητές (ή χαρακτηριστικά γνωρίσματα) της περιοχής $\mathbf{V} = \{x_i : i = 1, \ldots, n\}$, υποτίθεται ότι είναι ανεξάρτητες υπό όρους στην κρυμμένη μεταβλητή. Αυτό σκιαγραφείται στο σχήμα 2.7. Ο τρόπος με τον οποίο μια διαμόρφωση ανήκει σε μια ιδιαίτερη clustering μετριέται με τη βοήθεια της πιθανότητας.

Για αυτόν τον λόγο, αυτή η προσέγγιση-clustering έχει μια υγιή θεωρητική βάση.Όταν ένας σχηματισμός (configuration) εισάγεται ως στοιχείο στο ***naive Bayesian Network***, αποδίδει το συμπέρασμα (inference) παράγοντας τις πιθανότητες με τις οποίες αυτή η διαμόρφωση ανήκει στις διάφορες clusters, δηλ.τις καταστάσεις της κρυμμένης μεταβλητής. Δεδομένου ότι όλες οι πληροφορίες περιλαμβάνονται στις τιμές παραμέτρου (παρά στη δομή) του Μπαεζιανού δικτύου, η απεικόνιση των διάφορων clusters ή των scenarios μπορεί να είναι δύσκολη. Οι

διάφορες εφαρμόσιμες προσεγγίσεις για την εκμάθηση των παραμέτρων υπό την παρουσία μιας κρυμμένης μεταβλητής .

2.7 Structural Learning (Εκπαίδευση δομής)

Η δομική εκμάθηση, ή η επιλογή μοντέλου, ενδιαφέρεται για τον καθορισμό ενός Μπεϋζιανού δικτύου που περιγράφει τη κατανομή πιθανότητας που υπονοείται από τα δεδομένα μέχρι ενός ορισμένου βαθμού. Αυτός ο βαθμός μετριέται συνήθως με τη βοήθεια μιας αποκαλούμενης *συνάρτησης αποτελέσματος(**scoring function**)*. Το τελευταίο χαρτογραφεί, το ενδεχομένως υψηλό-διαστατικό, διάστημα των Μπαεζιανών δικτύων σε ένα μονοδιάστατο, χαρακτηριστικό στους πραγματικούς αριθμούς. Οι ιδιότητες των Μπαεζιανών δικτύων που περιγράφονται παραπάνω προτείνουν ορισμένα χαρακτηριστικά γνωρίσματα της συνάρτησης αποτελέσματος. Οι ιδιότητες των Μπαεζιανών δικτύων που περιγράφονται παραπάνω προτείνουν ορισμένα χαρακτηριστικά γνωρίσματα της συνάρτησης σκοραρίσματος .Οι κύριες συνέπειες αυτής της συσχέτισης μεταξύ των διαάφορων σχετικών αποτελεσμάτων***(scores***) είναι ότι ορισμένοι συνδυασμοί κατασκευασμένοι υπό όρους ανεξαρτησιών δεν μπορούν να συμπέσουν, και ότι μια επιτάχυνση των υπολογισμών μπορεί να επιτευχθεί. Πολλοί αλγόριθμοι εκμάθησης προσπαθούν να βρούν το (σφαιρικό) βέλτιστο της scoring function

Δυστυχώς, η αναζήτηση του βέλτιστου Μπαεζιανού δικτύου είναι ένα ***NP-hard*** πρόβλημα. Για αυτόν τον λόγο, κάποιος πρέπει να προσφύγει σε μια ευρετική *στρατηγική αναζήτησης* που μπορεί αποτελεσματικά να καθορίσει ένα Μπαεζιανό δίκτυο κοντά στο βέλτιστο.

Κεφάλαιο 3ο

3.1 Παρουσίαση δεδομένων

Σ ’αυτό το σημείο θα γίνει μια αναφορά στα διαθέσιμα δεδομένα και κατόπιν θα επιχειρηθεί μια πρώτη επαφή με τα προγράμματα σχετικά με τα Bayesian Netwoks.

Τα δεδομένα τα οποία θα μας βοηθήσουν στην κατασκευή του δικτύου καταγράφηκαν κατα τη διάρκεια μιας εργάσιμης ημέρας και αναφέρονται σε ελαττωματικά προιόντα(isdn modem).Πρόκειται για μια παρτίδα *600* μη λειτουργικών modem,ενώ τα παρατηρούμενα σφάλματα/cases είναι *956*.Ειναι σε ηλεκτρονική μορφή και σε xls(excel) format.Ένα μέρος των δεδομένων παρουσιάζονται στην ακόλουθη εικόνα

Α/Α	Σταθμός καταγραφής	Κωδικός ελαττώματος	Εξάρτημα	Πηγή
1	237	W01B	P08	402
1	910	W32F	M231	608
1	237	F08F	J87	209
2	910	W32F	M231	608
2	102	W302	L32	910
3	237	W02F	ST302	608
3	910	W32F	M231	608
4	910	W32F	M231	608
4	237	F08F	J87	209
5	910	W32F	ST300	608
6	910	K22F	T2	GT
7	910	W02F	M231	608
8	237	W02F	L27	608
8	910	W32F	M231	608
8	237	F08F	J87	209
9	910	W32F	M231	608
10	910	W32F	RY11	608
10	237	F08F	J87	209
11	101	S03A	M15	237
11	101	S03A	M16	237
11	910	W32F	M231	608
12	910	F08F	L11	608
12	237	W07B	N2	209
12	237	F08F	J87	209
12	237	F08F	DR2	209
13	910	F08F	L11	608
13	237	F08F	J87	209
13	237	F08F	DR2	209
14	910	F08F	L11	608
15	910	W02F	M232	608
15	101	W06F	M232	910
15	237	F08F	J87	209
15	237	K25B	KL2	GT
16	237	W02F	M99	608
16	910	W02F	M231	608
16	237	K25B	KL2	GT
17	910	F08F	L11	608
17	237	F08F	J87	209
18	237	W11B	C2	120
18	910	F08F	L11	608

Όπως μπορύμε να παρατηρήσουμε κάθε στήλη αντιπροσωπεύει μια μεταβλητή.Πιο συγκεκριμένα η στήλη με την ονομασία σταθμός καταγραφής(**QCS**)

αναφέρεται στους οπτικούς ελέγχους που πραγματοποιούνται κατά τη διάρκεια της παραγωγικής διαδικασίας.Όμως ήδη γνωρίζουμε από το πρώτο κεφάλαιο ότι στη γραμμή παραγωγής υπάρχουν τέσσερις οπτικοί έλεγχοι ,επομένως στη πρώτη στήλη καταγράφονται τέσσερις τιμές-οπτικοί έλεγχοι ***(237,910,101,102)***.

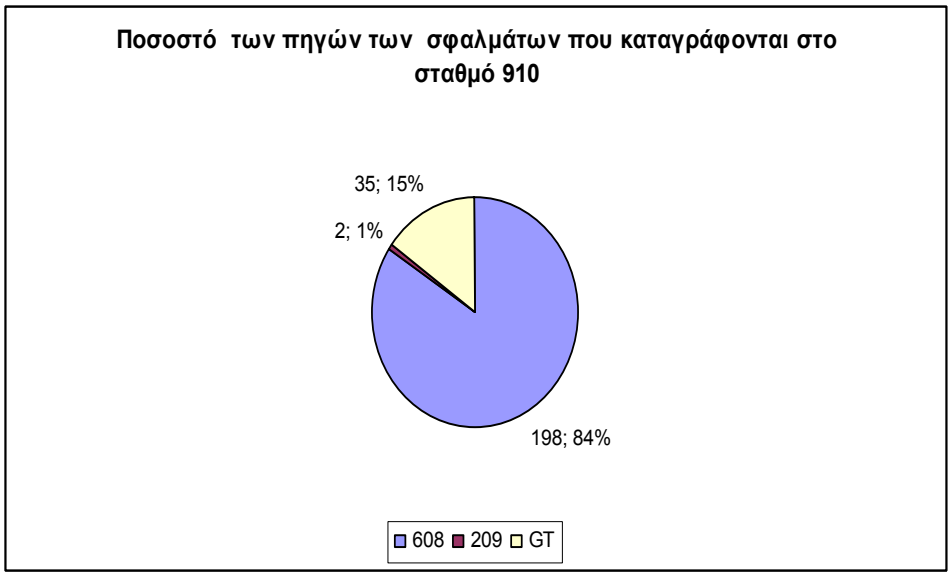
Αντίστοιχα η στήλη που εμφανίζει τον κωδικό σφάλματος (**code**) αναφέρεται στα παρατηρούμενα σφάλματα συνοδευόμενα με τον κωδικό τους για λόγους συντομίας και ευκολίας.Το πλήθος των σφαλμάτων φθάνει τον αριθμό των είκοσι πέντε:***(K04F,K22F,W02F,W04F,W06F,W32B,F02F,F04F,F08F,K03B,K11B,W01B,W03B,W05B,I31B,F03B,W31B,F01B,L31B,F21B,W32F,W11F,W302,WO7B,F11B).***

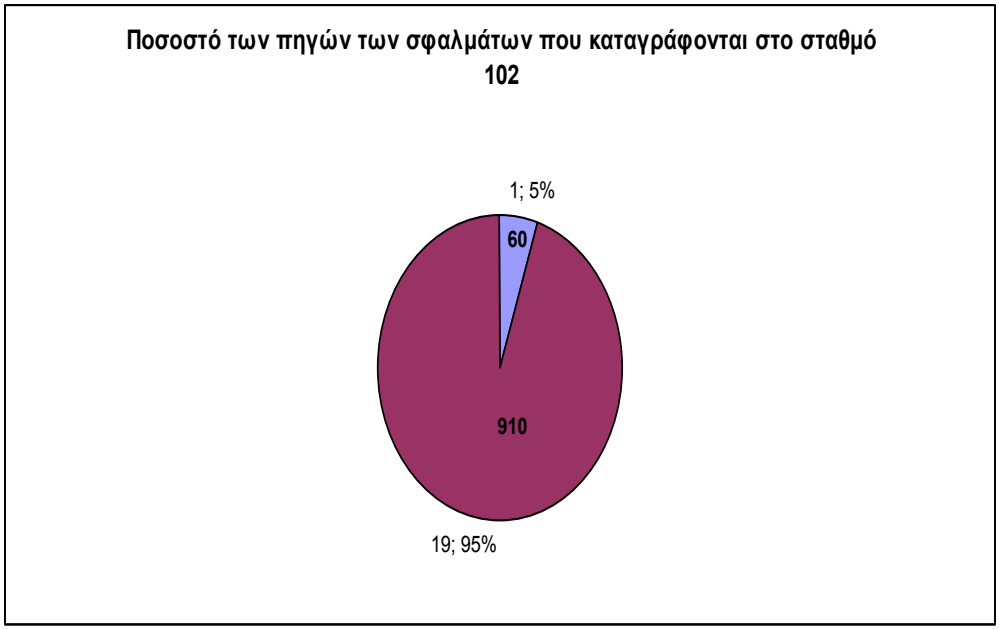
Αναφερόμενοι στη μεταβλητή **εξάρτημα** (***Device***) αυτή παίρνει 93 τιμές οι οποίες αντιστοιχούν στα χαλασμένα εξαρτήματα των προιόντων (**L33,DR1,P32**,....),ενώ οι τιμές της μεταβλητής **πηγή (**δηλαδή η αιτία που προκάλεσε το σφάλμα το οποίο καταγράφηκε από τον αντίστοιχο σταθμό) είναι 7 (**402,120,237,608,910,209,GT)**.

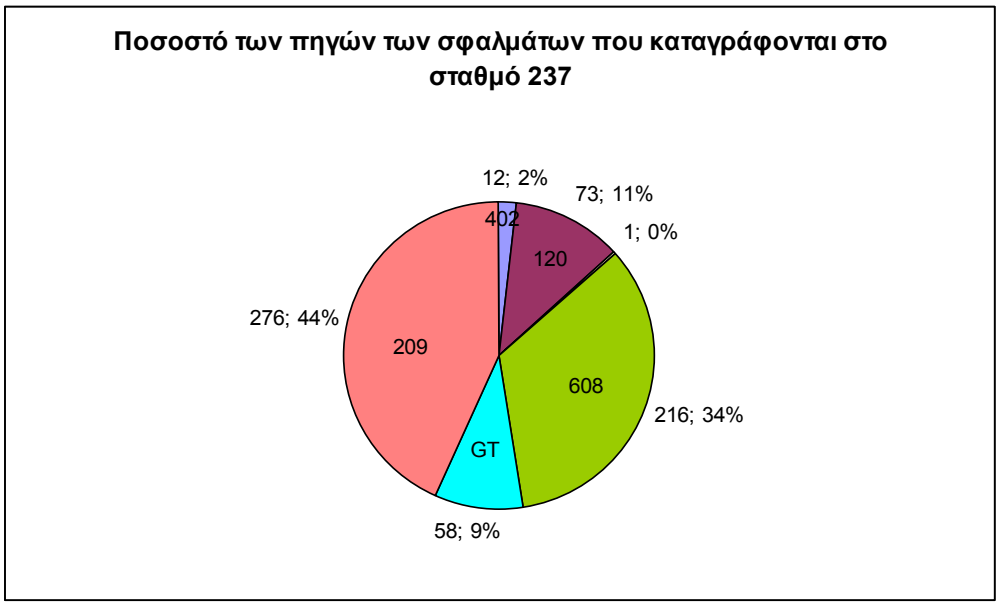
Τέλος η στήλη με την ονομασία ***A/A*** απλά αναφέρεται σε ποιο modem παρουσιάστηκε το/τα σφάλμα/σφάλματα και δεν λαμβάνονται υπόψη στη δημιουργία του δικτύου.

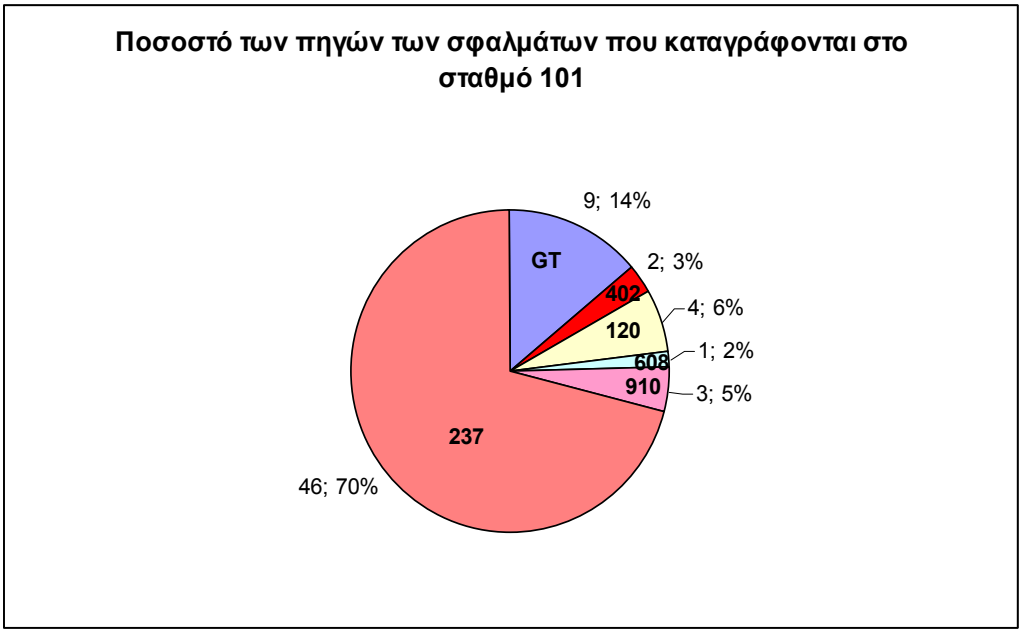
3.2 Στατιστική ανάλυση

Μια γρήγορη στατιστική ανάλυση των δεδομένων παρουσιάζεται στη συνέχεια και που συσχετίζει τις εσφαλμένες διαδικασίες με τους τέσσερις σταθμούς καταγραφής

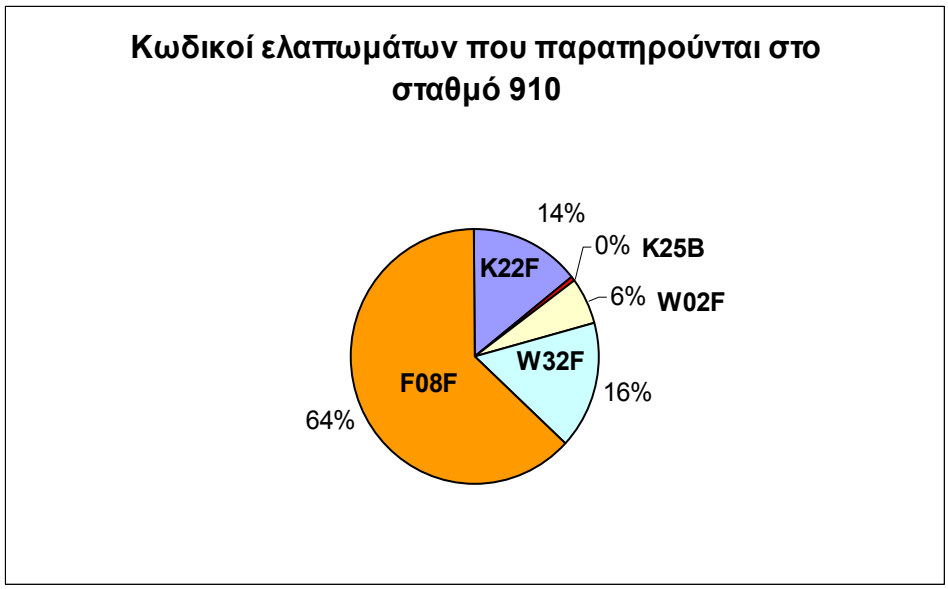



Ποσοστό των πηγών των σφαλμάτων που καταγράφονται στο σταθμό 102
1; 5%
60
910
19; 95%


Ποσοστό των πηγών των σφαλμάτων που καταγράφονται στο σταθμό 237
12; 2%
73; 11%
402
1; 0%
120
276; 44%
209
608
216; 34%
GT
58; 9%



Επίσης μπορούμε να απεικονίσουμε τη πιθανή εξάρτηση του σταθμού καταγραφής και των αντίστοιχων κωδικών σφάλματος.Για παράδειγμα ας δούμε ποιά σφάλματα παρατηρούνται στον σταθμό καταγραφής 910 και με τι ποσοστό:



Το επόμενο μέρος του παρόντος κεφαλαίου αναλώνεται στη σύντομη περιγραφή των προγραμμάτων που χρησιμοποιηθούν στη συνέχεια.

3.3 Παρουσίαση των λογισμικών πακέτων

3.3.1 HUGIN

Πρόκειται για ένα από τα πιο έγκριτα και διαδεδομένα προγράμματα στον χώρο των Bayesian Networks .Εμείς έχουμε στη διάθεσή μας μια εκπαιδευτική έκδοση πράγμα που μας δημιουργεί κάποιους περιορισμούς

Πρώτα αναπτύχθηκε στο πανεπιστήμιο του Aalborg της Δανίας με σκοπό την κατασκευή BBN(*Bayesian Belief Network*).Το Hugin Development Environment αποτελείται από τρία βασικά συστατικά/στοιχεία:Το Hugin Decision Engine,μια συλλογή από Application Program Interface και το Hugin Graphical User Interface.Το Hugin Graphical User Interface είναι ένα σημαντικό εργαλείο το οποίο παρέχει στο χρήστη τις δυνατότητες της Hugin Decision Engine.Βοηθάει στη κατασκευή μοντέλων τα οποία χρησιμοποιούνται σε διάφορες εφαρμογές. Επίσης το Hugin Graphical User Interface είναι ιδανικό για εκπαιδευτικούς σκοπούς.Για παράδειγμα όταν κάποιο άτομο πρωτοεισάγεται στον χώρο των Bayesian Network θα είναι προτιμητέο να μοντελοποιεί και να δοκιμάζει αυτά τα δίκτυα μέσω αυτού του πρακτικότατου εργαλείου.

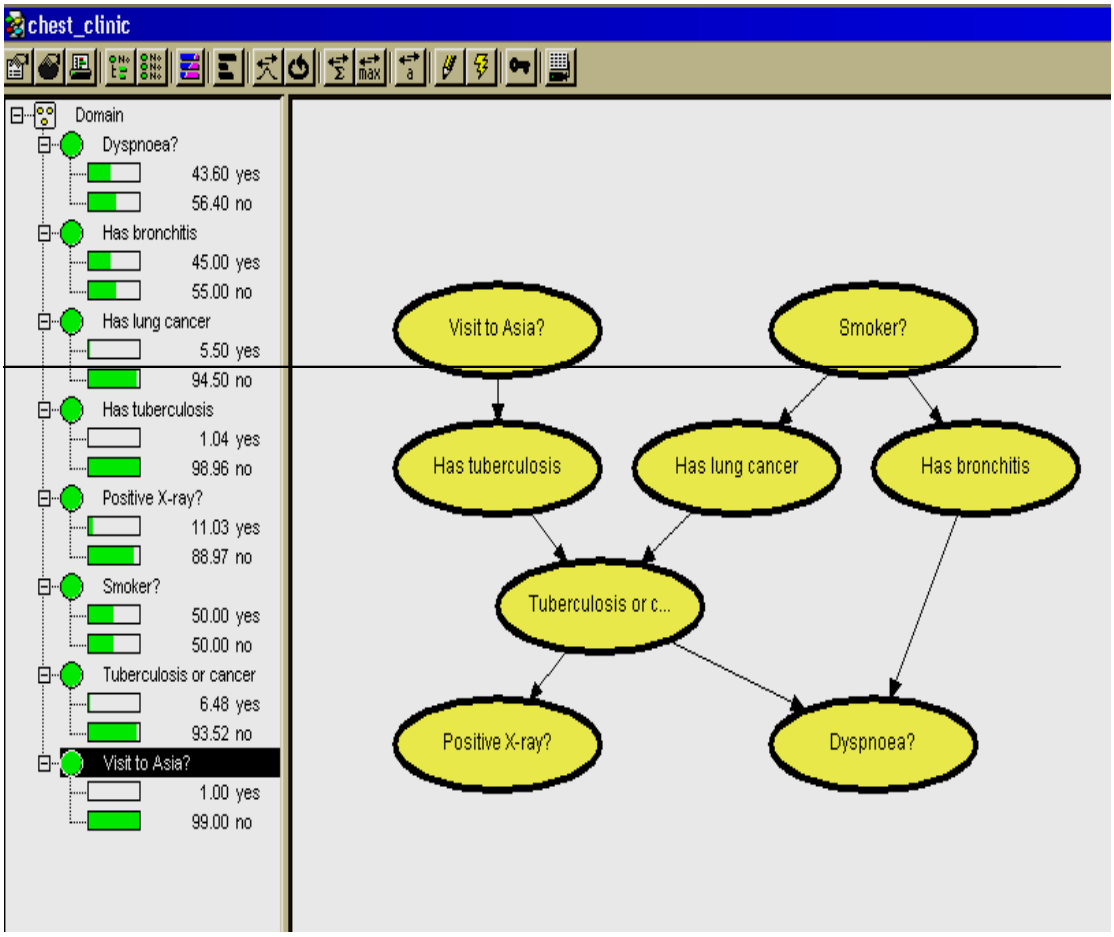
Οι βασικές λειτουργίες /δυνατότητες του είναι:

- παρέχει ένα σύνολο εργαλείων για την κατασκευή μοντέλων βασιζόμενων σε συστήματα υποστήριξης αποφάσεων σε domains χαρακτηριζόμενα από έμφυτη αβεβαιότητα.Αύτα είναι τα Bayesian Networks(**BNs**) και η επέκτασή τους με μεταβλητές απόφασης(**decision node**) και μεταβλητές χρησιμότητας(**utility nodes**) τα διαγράμματα απόφασης(**influence diagrams**)
- υποστηρίζει διακριτές (**discrete**)καθώς επίσης και συνεχείς μεταβλητές(**continuous**)

●παρέχει ένα ισχυρότατο μηχανισμό για τη δημιουργία μοντέλων με αλλεπάληλα πρότυπα ,ενώ επίσης κατασκευάζει μοντέλα σε ιεραρχική δομή(**object-oriented networks**)
●κατασκευάζει αυτόματα το δίκτυο υποστηρίζοντας την εισαγωγή δεδομένων με την χρήση των αλγόριθμων **PC** και **NPC** algorithm (**structure learning**)
● με ανάλογη διαδικασία (μέσω δεδομένων) μπορούν οι πιθανότητες των μεταβλητών του δικτύου να γίνουν γνωστές αυτόματα όταν μόνο η δομή του μοντέλου ειναι γνωστή(**EM learning**)
● υποστηρίζει την αυτόματη ανανέωση των πιθανοτήτων των μεταβλητών του απαρτίζουν το δίκτυο λαμβάνοντας νέα στοιχεία.Ο συγκεκριμένος αλγόριθμος λειτουργεί μόνο με διακριτές μεταβλητές (**adaptation-sequential updating**).

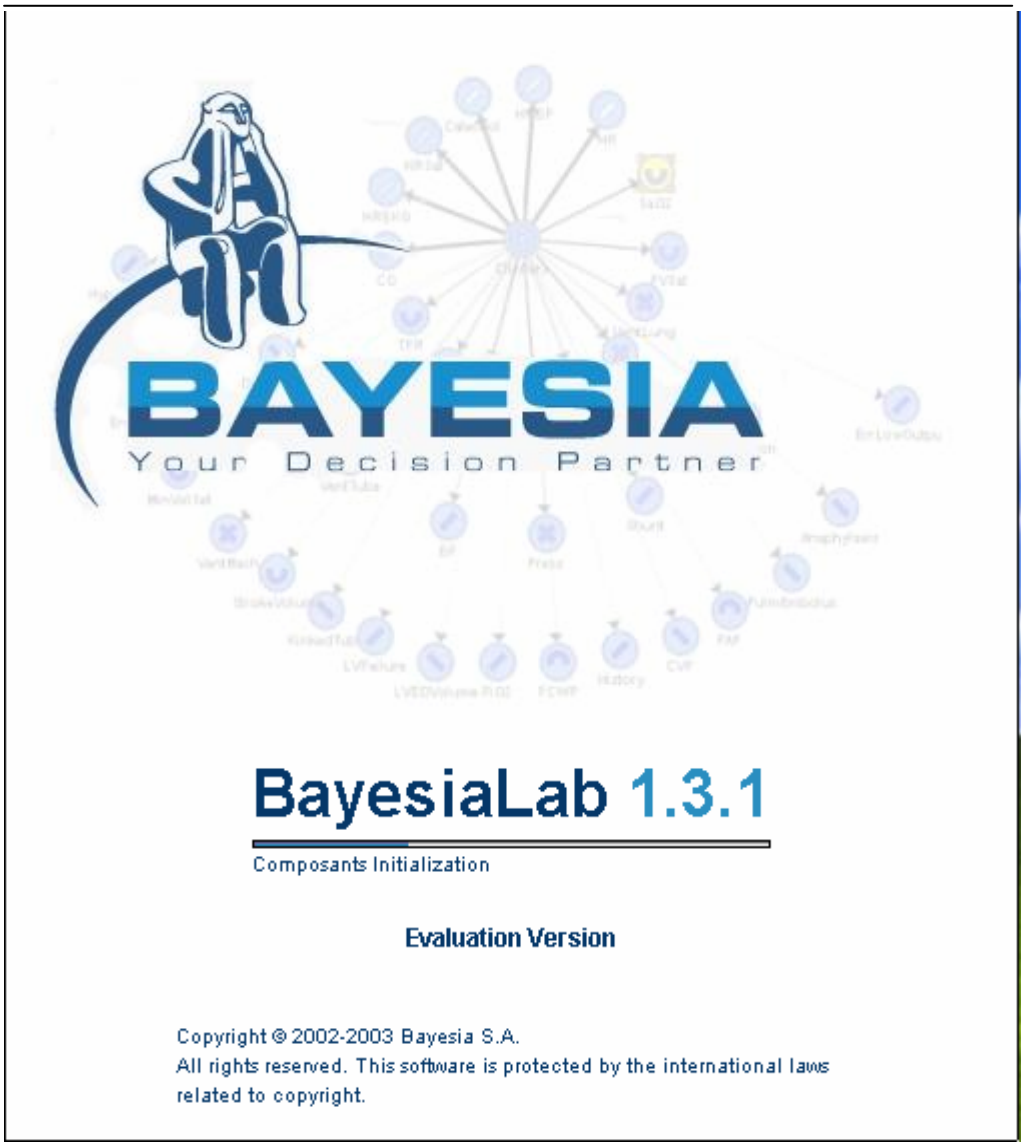
Το Hugin είναι διαθέσιμο σε C,C++ και Java «βιβλιοθήκες»(libraries) όπως και σε ActiveX server.

Ένα δικτύο δημιουργημένο στο Hugin έχει αυτή την εικόνα:



Περισσότερες πληροφορίες υπάρχουν στο site της εταιρείας: ***http://www.hugin.com***

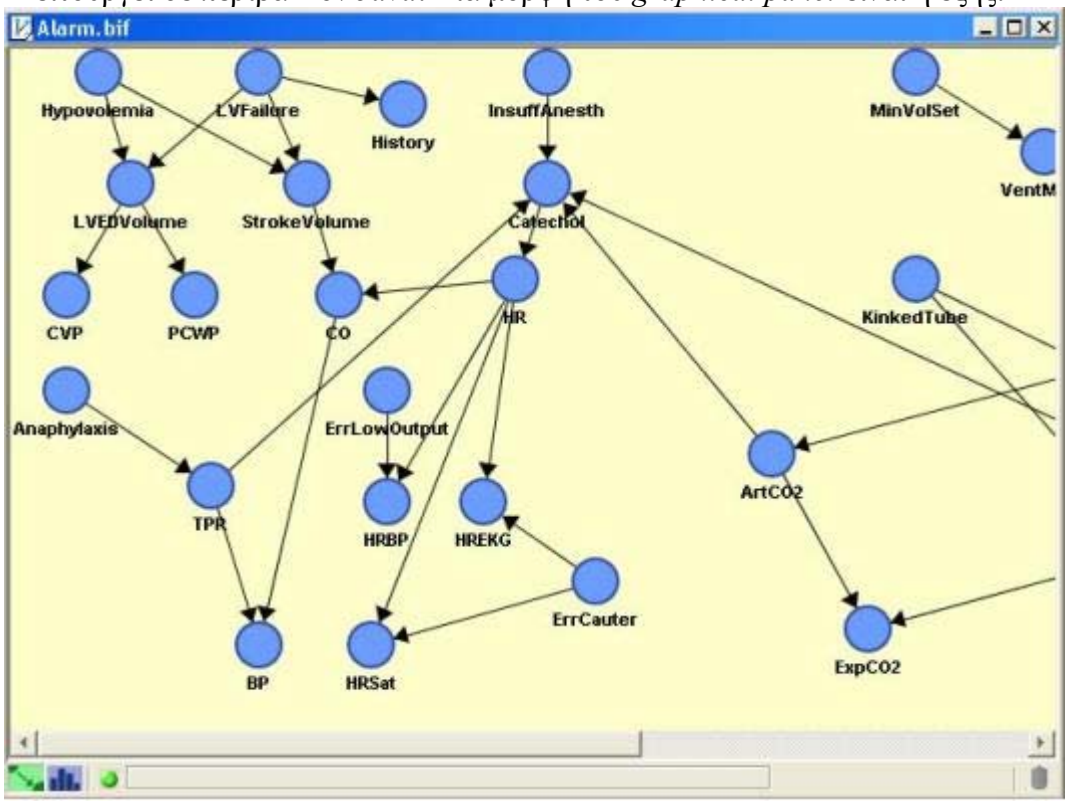
BAYESIALAB



Επίσης πρόκειται για ένα από τα δημοφιλέστερα ,αναγνωρίσιμα και αξιόπιστα προγράμμα που βρίσκονται στο εμπόριο.Έχει κατασκευαστεί από γάλλους επιστήμονες. Το Bayesialab επιτρέπει τον γραφικό έλεγχο των Bayesian Networks.Μπορεί να διορθώνει, να κατασκευάζει,να ορίζει, να «εκπαιδεύει» τα μοντέλα.Για περισσότερες λεπτομέρειες στο site: ***www.bayesia.com***

Βέβαιως όπως και προηγουμένως διαθέτουμε την εκπαιδευτική του έκδοση η οποία εμπεριέχει κάποιους περιορισμούς.

Λειτουργεί σε περιβάλλον *Java*.Μια μορφή του *graphical panel* είναι η εξής:



● Επιτρέπει τον αυτόματο σχεδιασμό του δικτύο εισάγοντας μια βάση δεδομένων
●Υποστηρίζει την ανανέωση της βάσης δεδομένων πάντα σε συμφωνία με τον πιθανοτικό νόμο που περιγράφει το Bayesian Network
●Υποστηρίζει τόσο τις διακριτές μεταβλητές ,όσο και τις συνεχείς καθώς επίσης και τις καθορισμένες μεταβλητές(δηλαδή οι σχέσεις τους με τους γονείς δεν είναι πιθανοτική αλλά λογική)
●Διαθέτει ένα σύνολο εκπαιδευτικών αλγορίθμων από αλγόριθμους παραμετρικής εκπαίδευσης (εκτιμάει τις κατα συνθήκη πιθανότητες βασιζόμενο στη συχνότητα εμφάνισης των μεταβλητών της βάσης δεδομένων),εώς αλγόριθμους δομικής εκπαίδευσης(**structural learning algorithms**).
● Παρέχει ένα ολοκληρωμένο σύστημα ανάλυσης των Bayesian Networks.Διακρίνονται σε τρία επίπεδα:
1.Ανάλυση της δύναμης των τόξων(έλεγχος των σχέσεων μεταξύ των συσχετισμένων μεταβλητών)
2. Ανάλυση των σχέσεων μεταξύ της «κεντρικής» μεταβλητής(**target node**) και των υπόλοιπων
3. Τοπική ανάλυση του δεσμού ανάμεσα των μεταβλητών και των καταστάσεων της «κεντρικής» μεταβλητής
● Υπολογίζει την ποιότητα του υπάρχοντος δικτύου για πρόβλεψη της «κεντρικής» μεταβλητής με χρήση της βάσης δεδομένων
●Περιέχει και εργαλεία για τα Dynamics Bayesian Networks

●Μέσω ενός συνόλου αλγορίθμων επιτρέπει την ανεπιτήρητη εκπαίδευση (**unsupervised learning**) όλων των πιθανοτικών σχέσεων που υπάρχουν μέσα στη βάση
Τρεις μέθοδοι έρευνας με διαφορετική λογική προτείνονται:
ο **SopLEQ**:στηρίζεται στον ολικό χαρακτηρισμό της βάσης δεδομένων και την αξιοποίηση των ισοδύναμων ιδιοτήτων των Bayesian δικτύων(γρήγορη)
ο**Taboo :**δομική εκμάθηση (structural learning) που βελτιώνει τη Taboo έρευνα
ο **Taboo order:**κάνει χρήση της Taboo έρευνας για να βρει τη βέλτιστη εντολή των κόμβων(ο πιο ολοκληρωμένος άρα και ο πιο χρονικά μεγάλος)

● Ένα από τα μενού που διαθέτει επιτρέπει σε κάποιους αλγόριθμους να κάνουν κατηγοριοποίηση των δεδομένων με ένα τρόπο ανεπιτήρητο με στόχο να βρούμε μέρη με ομογεννή στοιχεία.
●μέσω ενός συνόλου αλγορίθμων επιτρέπει την εκπαίδευση του δικτύου έχοντας συγκεκριμένη δομή,βασισμένη γύρω από την «κεντρική» μεταβλητή
Εν συντομία αυτοί οι αλγόριθμοι είναι:
●**Naïve architecture**
●**Augmented Naïve A rchitecture**
●**Sons & spouses learning**
● **Markov Blanket learning**
● **Augmented Markov Blanket learning**

Αυτές είναι οι σημαντικότερες λειτουργίες των προγραμμάτων ,των οποίων οι δυνατότητες δεν είναι δυνατό να περιγραφούν επακριβώς στα πλαίσια του παρόντος κεφαλαίου.

Κεφάλαιο 4ο

4.1 Εισαγωγή

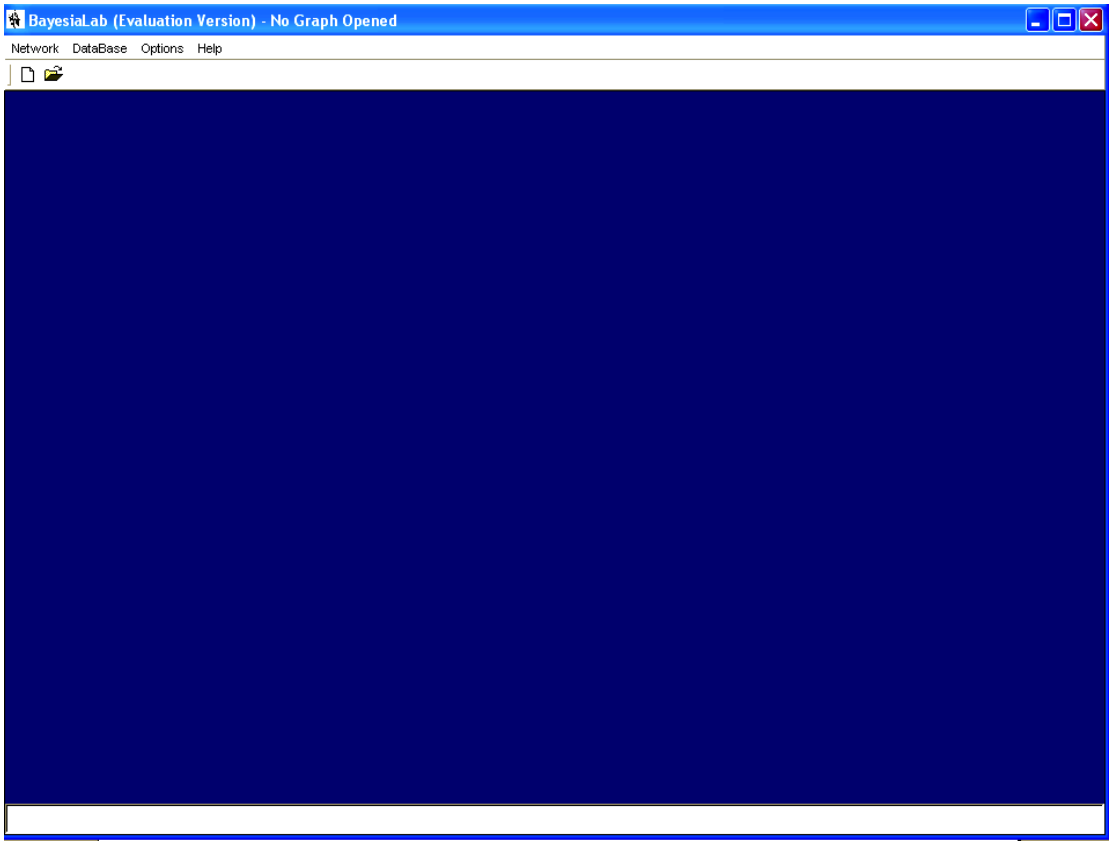
Πρωτού ξεκινήσουμε με την περιγραφή της εξαγωγής των δικτύων μέσω των δίαφορων λειτουργιών των προγραμμάτων ,θα ήταν λογικό να υπενθυμίσουμε ότι ο αριθμός των μεταβλητών του Μπαεζιανού δικτύου φθάνει τον 129, οι οποίες κατηγοριοποιούνται σε τέσσερις κύριες μεταβλητές όπου καταγράφονται ως τιμές-καταστάσεις της επιμέρους μεταβλητής.Αυτές οι μεταβλητές είναι οι ακόλουθες:

- ***QCS*** (αναφέρεται στους σταθμούς καταγραφής των σφαλμάτων-4 καταστάσεις)
- ***CODE*** (έχει ως καταστάσεις τους κωδικούς σφαλμάτων-25 καταστάσεις)
- ***DEVICE***(δηλώνει τα εξαρτήματα όπου παρουσιάστηκαν τα σφάλματα-93 καταστάσεις)
- ***SOURCE***(αναφέρεται στην αιτία που προκάλεσε το σφάλμα-7 καταστάσεις)

Τέλος αξίζει να σημειωθεί ότι ενώ οι δυνατοί τρόποι κατασκευής ενός Bayesian δικτύου είναι δύο(είτε manually ,είτε μέσω εισαγωγής δεδομένων-structure learning). Στην περιπτωσή μας θα προτιμηθεί ο δεύτερος τρόπος,μια και ο «χειρονακτικός» τρόπος χρησιμοποιείται από τους ειδικούς στον χώρο.

4.2 Εφαρμογή του Bayesialab στη δημιουργία του δικτύου

Η αρχική μορφή του προγράμματος είναι:

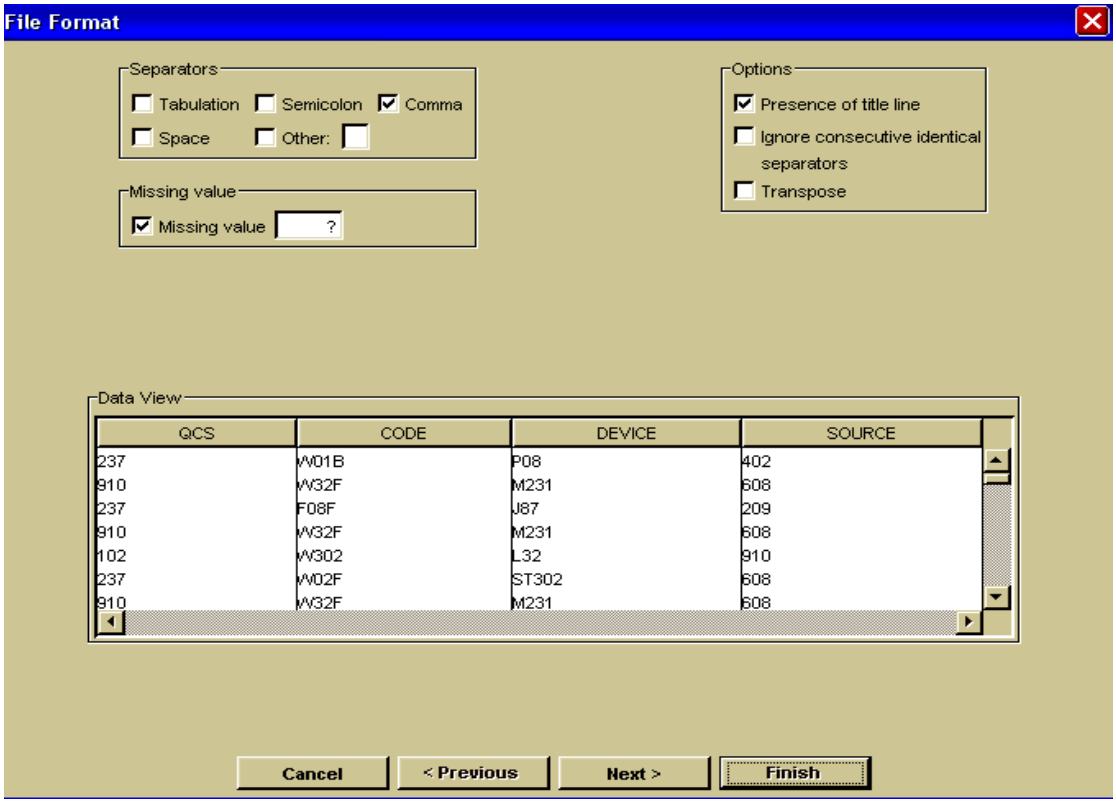


4.2.1 Εισαγωγή δεδομένων

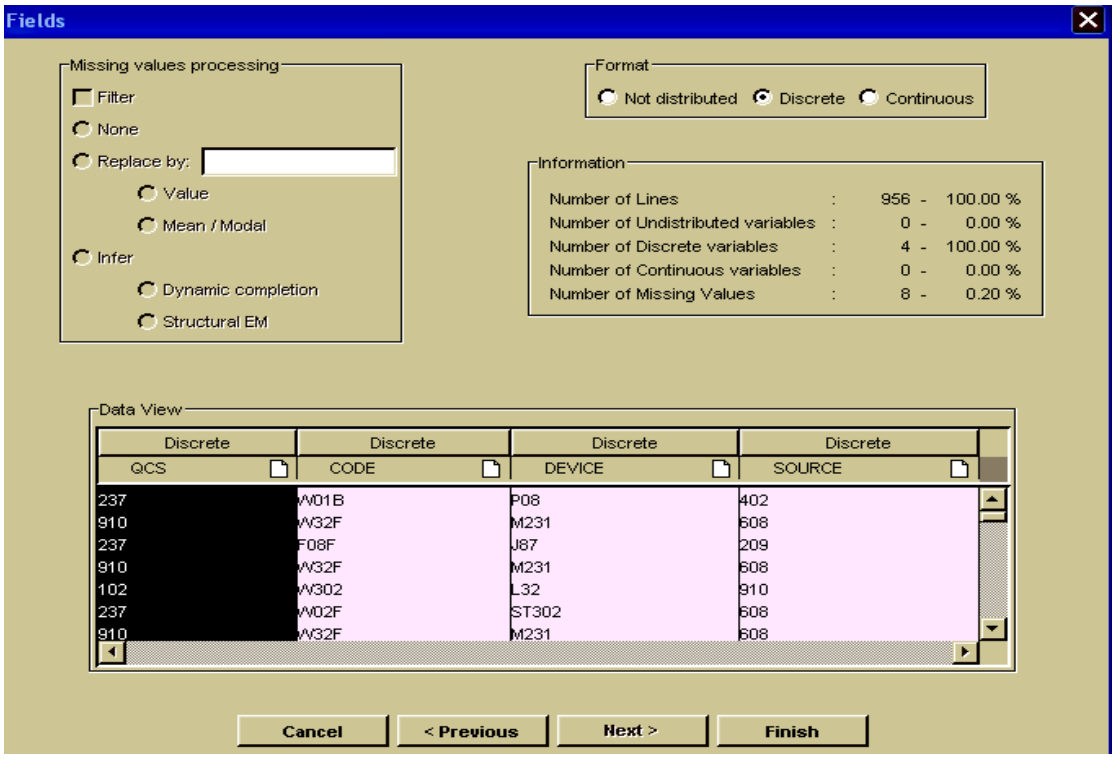
Το επόμενο βήμα είναι η επιλογή της κατάλληλης βάσης δεδομένων μέσω της επιλογής ″ ***open database*** ″ που υπάρχει στο ***command zone***.

Αφού έχουμε επιλέξει ποια βάση δεδομένων θα επεξεργαστούμε (ένα txt αρχείο) ,στη συνέχεια θα εμφανιστεί ένα μενού που μας βοηθάει στον προσδιορισμό της μορφής των δεδομένων που θα χρησιμοποιηθούν στη συνέχεια.Ειδικότερα μπορούμε να επιλέξουμε με ποιους χαρακτήρες θα γίνει ο διαχωρισμός ,αν υπάρχει κάποια γραμμή που να εμφανίζεται στην αρχή των δεδομένων και να περιέχει τα ονόματα

των μεταβλητών ,ενώ στην περίπτωση που δεν έχουμε καταγραφή των καταστάσεων των μεταβλητών με ποιο τρόπο θα δηλώνουμε αυτή την απουσία:



Στη συνέχεια εμφανίζεται ένα δεύτερο βοηθητικό μενού ,το οποίο επιτρέπει τον προσδιορισμό του είδους των μεταβλητών (διακριτή ή συνεχής)



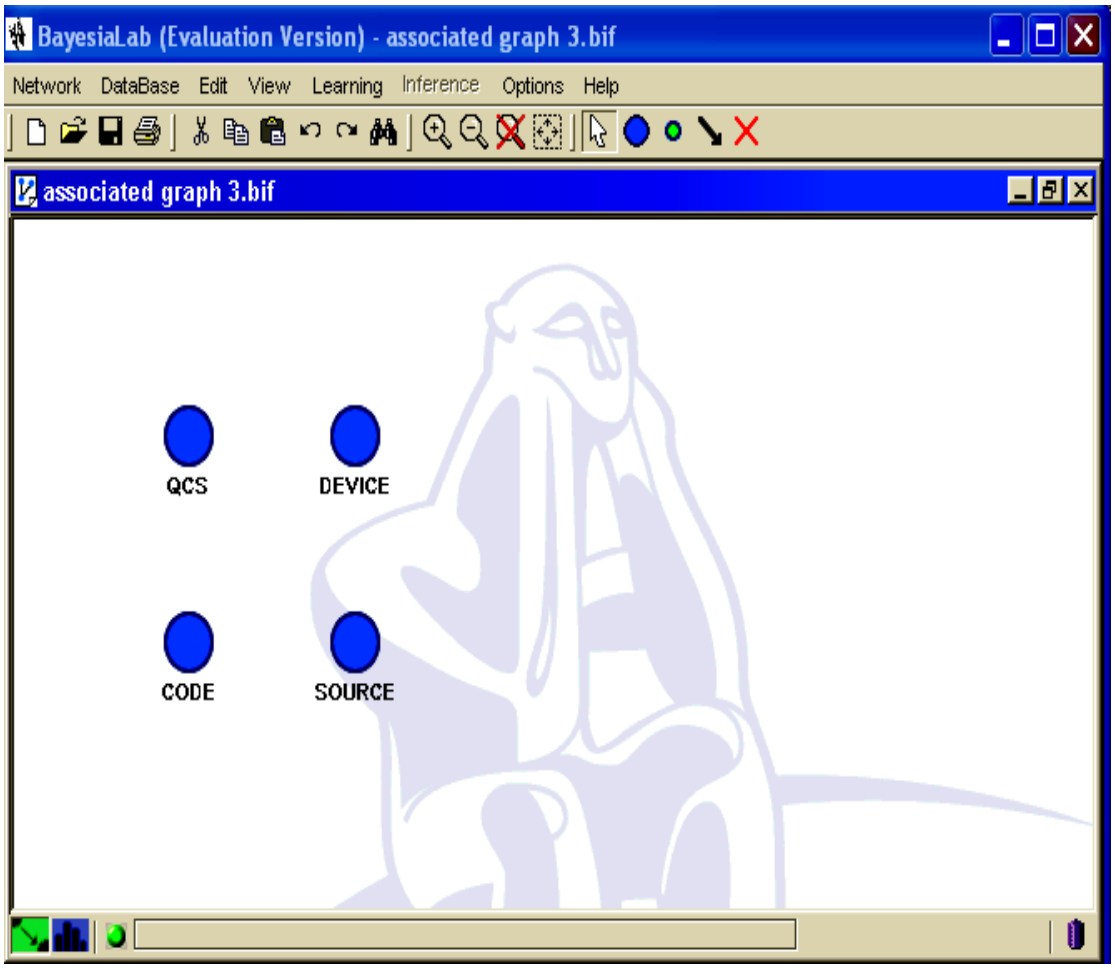
Αυτό το μενού μας παρέχει πληροφορίες σχετικές με τη δομή των δεδομένων:

Information

Number of Lines	:	956 -	100.00 %
Number of Undistributed variables	:	0 -	0.00 %
Number of Discrete variables	:	4 -	100.00 %
Number of Continuous variables	:	0 -	0.00 %
Number of Missing Values	:	8 -	0.20 %

Όπως φαίνεται ,η βάση των δεδομένων αποτελείται από 956 cases,ο αριθμός των διακριτών μεταβλητών είναι 4 ,ενώ υπάρχουν και 8 μη διαθέσιμες καταστάσεις.

Όταν το δεύτερο μενού κλείσει ,ένα "φύλλο εργασίας'' εμφανίζεται (worksheet):

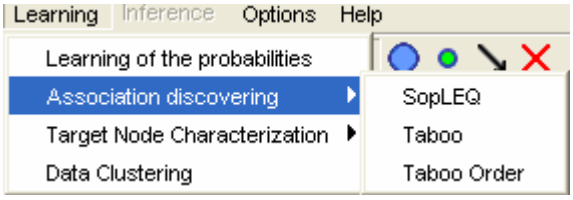


4.2.2 Επιλογή *unsupervised* αλγόριθμου

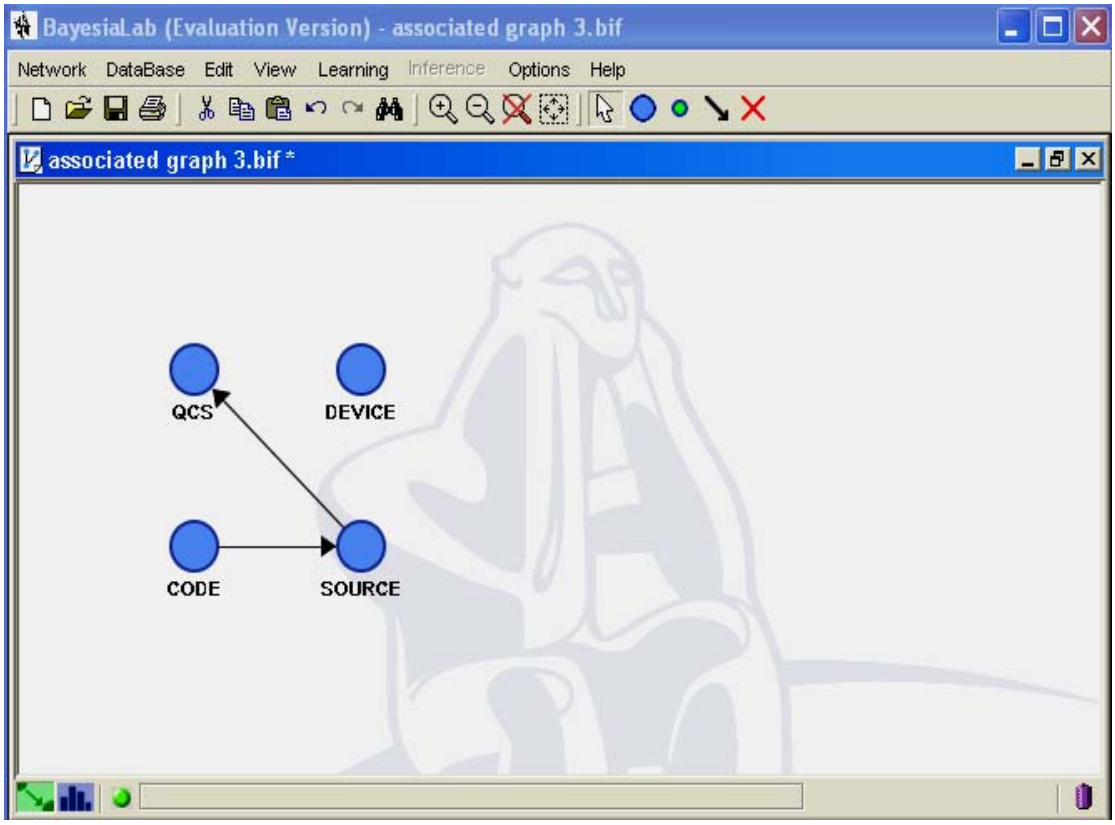
Η επόμενη κίνηση είναι η επιλογή των αλγόριθμων εκπαίδευσης για να μορφοποιηθεί το δίκτυο. Το Bayesialab χρησιμοποιεί τρεις αλγόριθμους ‘΄βοηθητικής ανακάλυψης’’(association discovery):

ο**SopLEQ**:στηρίζεται στον ολικό χαρακτηρισμό της βάσης δεδομένων και την αξιοποίηση των ισοδύναμων ιδιοτήτων των Bayesian δικτύων(γρήγορη)
ο**Taboo :**δομική εκμάθηση (structural learning) που βελτιώνει τη Taboo έρευνα
ο**Taboo order:**κάνει χρήση της Taboo έρευνας για να βρει τη βέλτιστη εντολή των κόμβων(ο πιο ολοκληρωμένος άρα και ο πιο χρονικά μεγάλος)

Εμείς θα κάνουμε χρήση του **Taboo order** λόγω της πιο ολοκληρωμένης μορφής του αλγόριθμου με συνέπεια την αξιόπιστη απεικόνιση της φύσης του δικτύου.

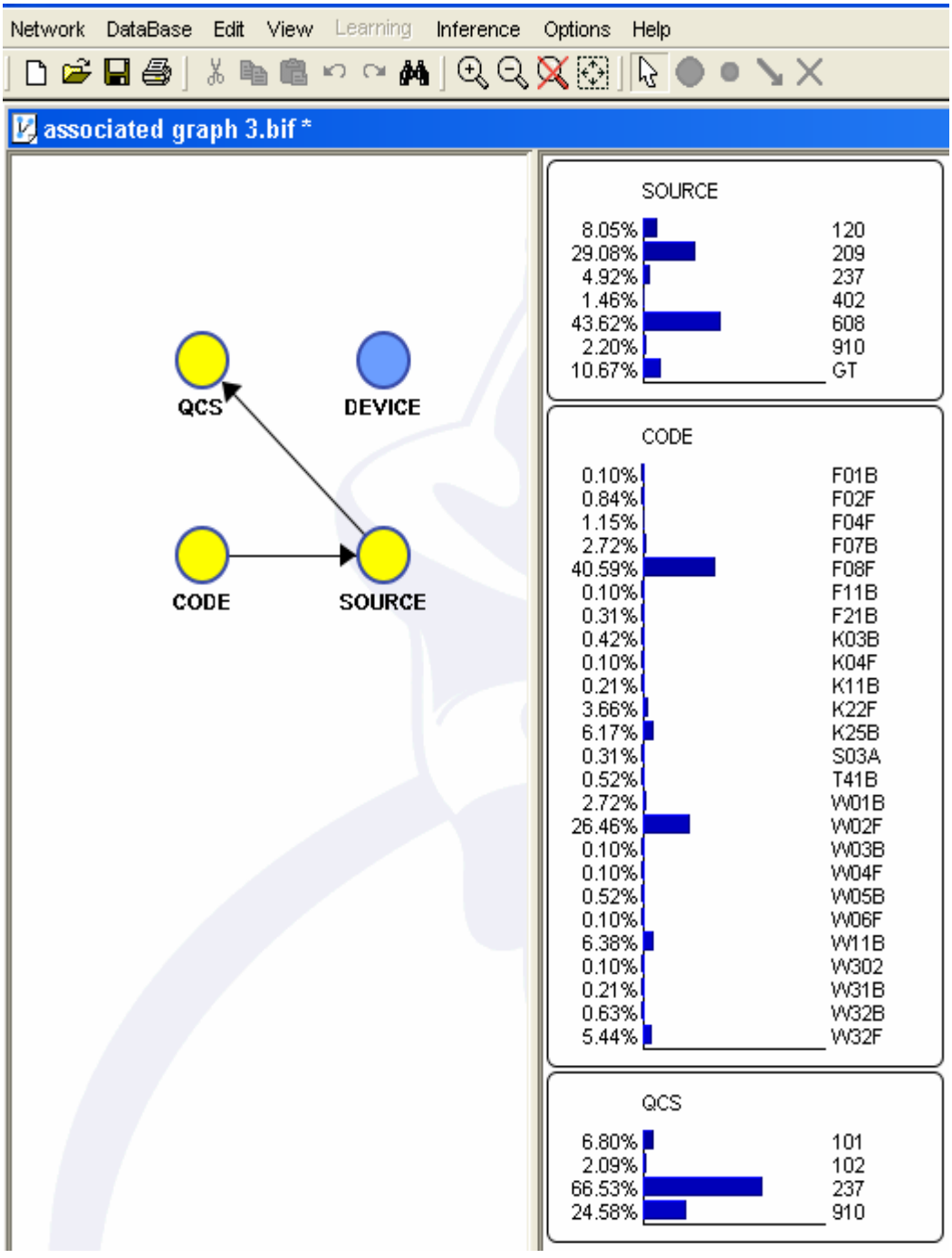


Αυτό που δημιουργήθηκε έχει την ακόλουθη μορφή:

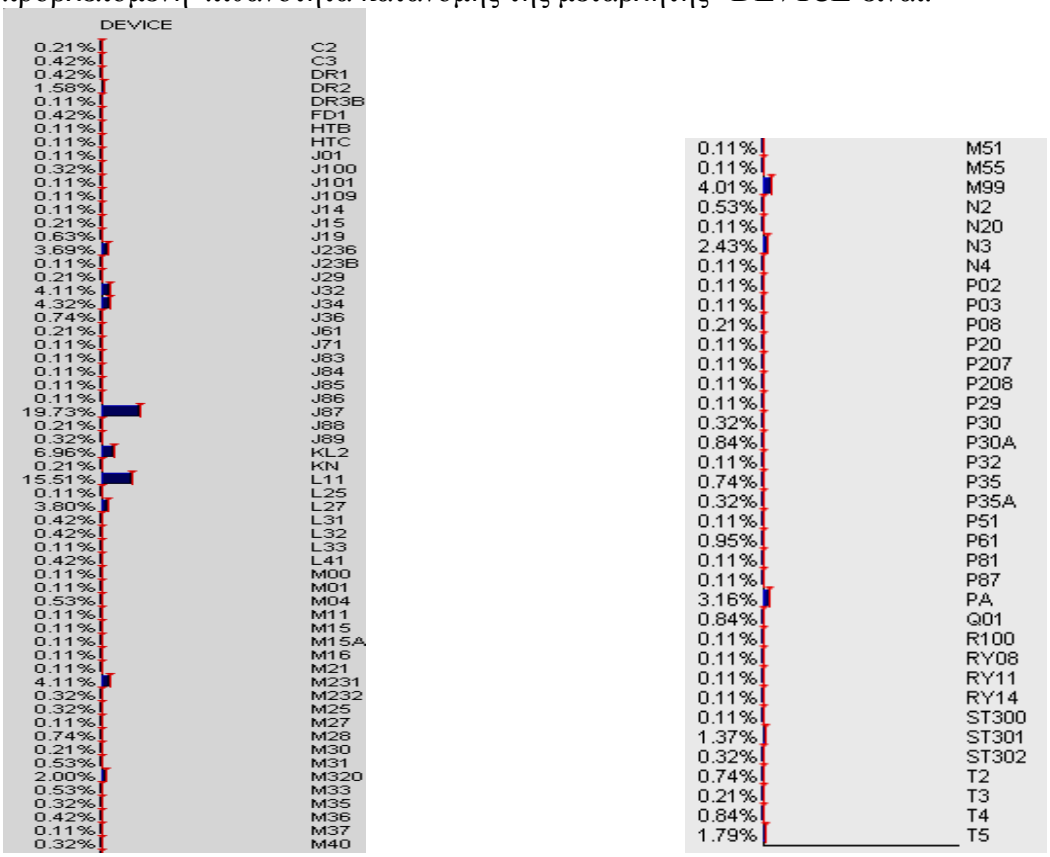


4.2.3 Monitoring(Απεικόνιση κατανομών)

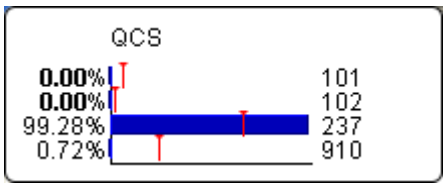
Τώρα θα δοθεί μια αναλυτικότερη εικόνα των μεταβλητών(πλην αυτής της ***CODE***) του δικτύου:

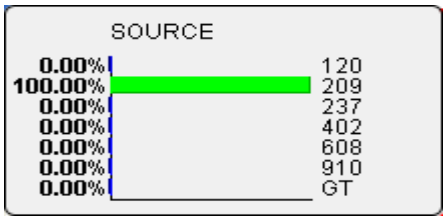


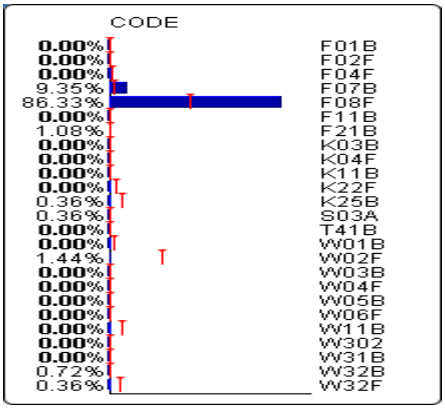
Η προβλεπόμενη πιθανότητα κατανομής της μεταβλητής ***DEVICE*** είναι:



Παρατηρούμε ότι η κατάσταση ***209*** έχει καταγραφεί σε ποσοστό 29,08%.Έστω ότι θέλαμε να δούμε τι συνέπειες θα προκαλούνταν στο δίκτυο σε περίπτωση που η εσφαλμένη διαδικασία θα ήταν σε ποσοστό 100% η κατάσταση ***209.***Αυτό που έχουμε να κάνουμε είναι απλά ένα διπλό κλικ πάνω στη κατάσταση ***209*** και μετά να αναλύσουμε τις μεταβλητές.



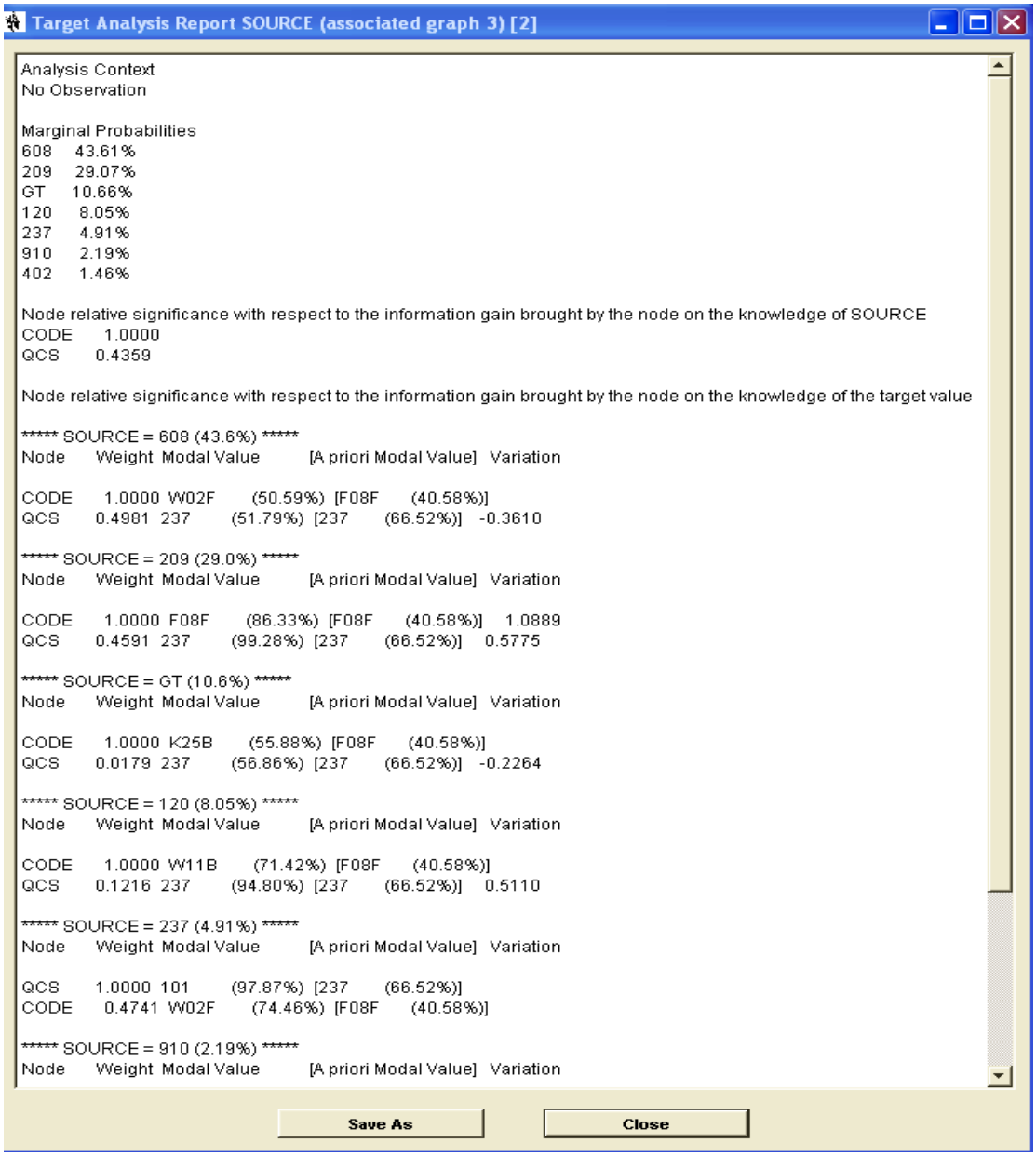




Λόγω της διάδοσης (**propagation**) της νέας πληροφορίας ,οι κατανομές πιθανότητας για κάθε μεταβλητή έχει ανανεωθεί.Η πιο αξιοσημείωτη αλλαγή που παρατηρούμε είναι ότι ο σταθμός **237** είναι αυτος που θα καταγράφει το σφάλμα της **209** σε ποσοστό **99,28%**,ενώ καταγράφεται μια αύξηση του κωδικού σφάλματος ***F08F*** (από 40,59% σε 86,33%).Οι μεταβολές μεταξύ των τιμών των μεταβλητών διακρίνονται μέσω του μικρού κόκκινου συμβόλου(δεν υπάρχει απεικόνιση της DEVICE λόγω της πολύ μικρής μεταβολής της κατανομής της πιθανότητας).

4.2.4 Target analysis Report

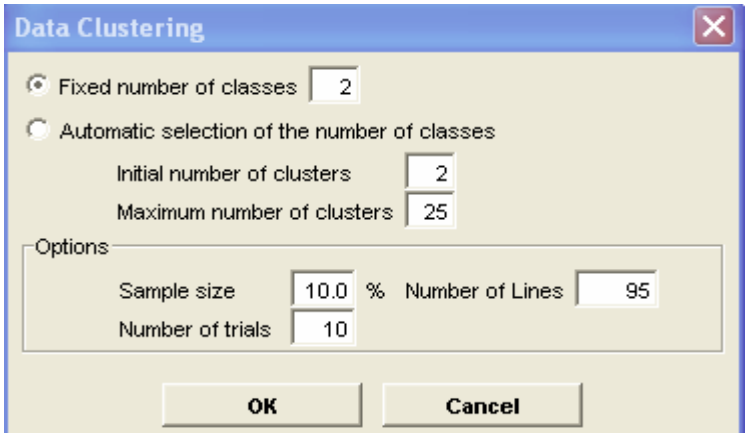
Το **Bayesialab** παρέχει στον χρήστη μια αναλυτική αναφορά και πληροφορίες για την κεντρική μεταβλητή (target node).Αυτό πραγματοποιείται θέτοντας την μεταβλητή **SOURCE** ως target node και κατόπιν επιλέγοντας την εντολή :**Target analysis Report** μέσα στο **inference menu.**



Αυτή η αναφορά περιέχει:το σύνολο των παρατηρήσεων(analysis context), την οριακή κατανομή(***marginal distribution***) της target node ,μια λίστα των μεταβλητών κατατεγμένες ανάλογα με το μέγεθος της πληροφορίας που παρέχουν στη γνώση της target node και των επιμέρους καταστάσεών της.

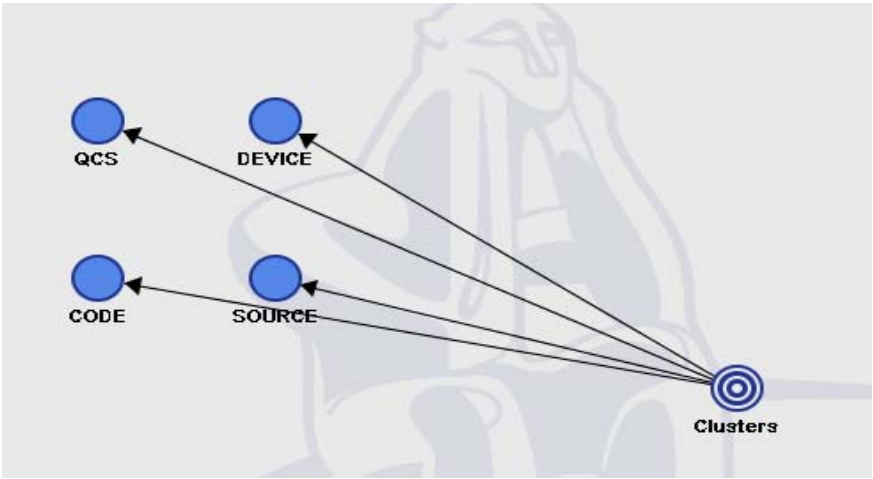
4.2.5 Αλγόριθμοι κατηγοριοποίησης (Clustering)

Επιπροσθέτως με την *association discovery* ,το Bayesialab διαθέτει ***clustering*** αλγόριθμους.Ο στόχος αυτών των μεταβλητών είναι η ανακάλυψη «φυσικών» διαχωρισμών από τις παρατηρήσεις που καταγράφονται στη βάση δεδομένων.Αυτοί οι διαχωρισμοί μοιράζονται ένα σύνολο κοινών ιδιοτήτων.Το ***learning menu*** μας επιτρέπει την εμφάνιση του ***clustering assistant***

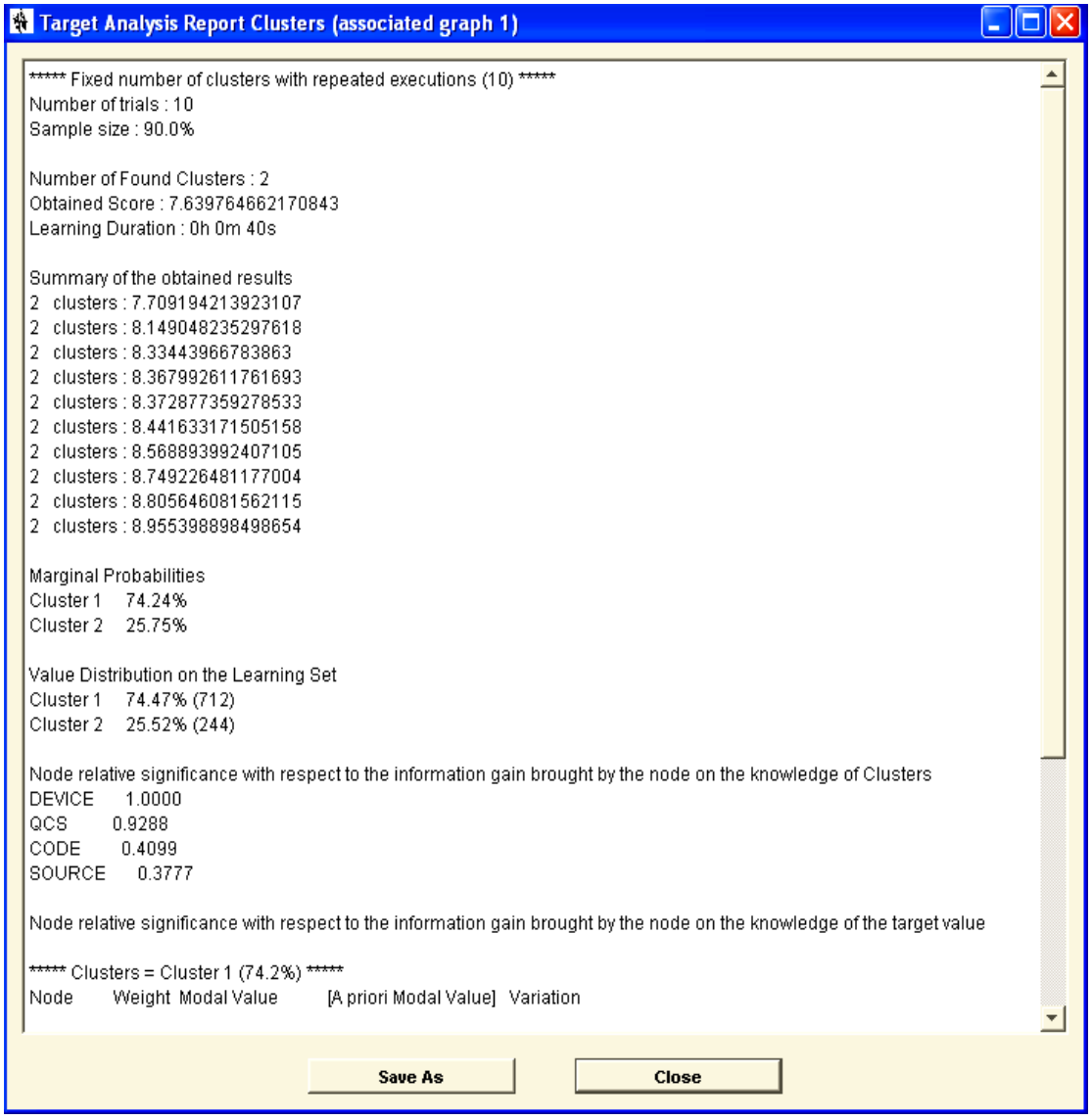


Οι προτεινόμενες μέθοδοι είναι δύο:

- **Fixed number of classes:**ο αριθμός των κατηγοριών καθορίζεται από τον χρήστη
- **Automatic selection of the number of classes**:το σύστημα ψάχνει από μόνο του τον βέλτιστο αριθμό των κατηγοριών (clusters) για τον όσο τον δυνατό καλύτερο διαχωρισμό.Αυτή η μέθοδος στηρίζεται σε ένα **directed random walk.**Αρχικά πρέπει να οριστεί ο αρχικός αριθμός των κατηγοριών ,καθώς και ένας μέγιστος αριθμός τους.Επίσης μπορεί να ρυθμιστεί ο αριθμός των βημάτων του random walk,όπως και το μέγεθος του δείγματος που θα χρησιμοποιηθεί για την εκπαίδευση του δικτύου.Όσον αφορά τον αριθμό των βημάτων ,αυτός μπορεί να τεθεί πολύ μεγάλος με το σκεπτικό ότι η διαδικασία σταματάει όποτε το επιθυμούμε εμείς.



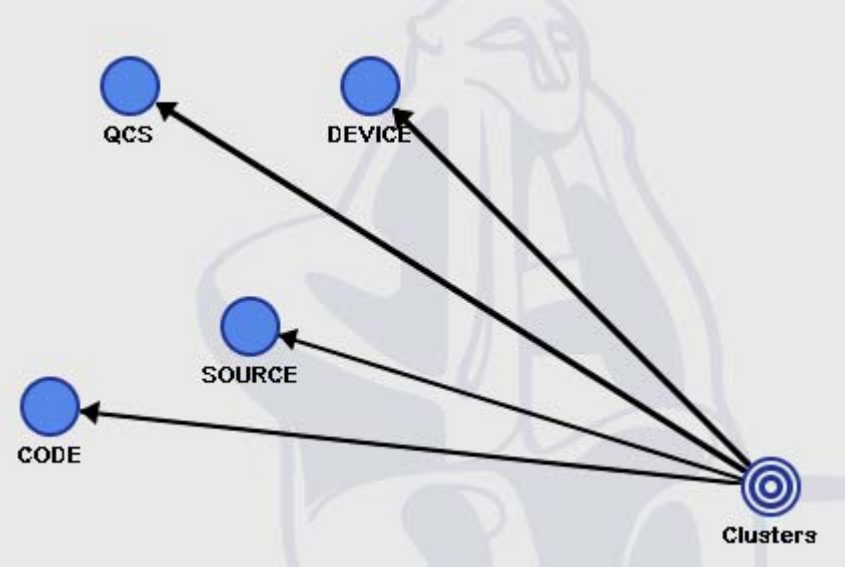
Η επιλογή μας ήταν η **Fixed number of classes** με ορισμένη τιμή των clusters 2 και στο τέλος του αλγόριθμου εκπαίδευσης εμφανίζεται μια ανάλογη αναφορά με την προηγούμενη.Περιέχει τους παράμετρους της εκπαίδευσης αλλά και τα αποτελέσματα,δίνει τις οριακές κατανομές των κατηγοριών που καταγράφηκαν,ταξινομημένες ανάλογα με το «βάρος» τους αλλά και την κατανομή των ***clusters*** στην βάση δεδομένων η οποία χρησιμοποιήθηκε για την εκπαίδευση.



Παρατηρούμε ότι ο τελικός αριθμός των clusters είναι 2 ,το αποτέλεσμα (score) του αλγόριθμου είναι 7,63976 και επίσης υπάρχει ταξινόμηση των μετβλητών ανάλογα με την πληροφορία που στέλνει στη cluster node (δηλαδή κατάταξη με κριτήριο την σημαντικότητα της μεταβλητής ως προς την cluster node).Η σείρα κατάταξης είναι πρώτα η **DEVICE(1.000),QCS(0,9288),CODE(0,4099)** και **SOURCE(0,3777).**Στο τελευταίο μέρος της αναφοράς υπάρχει μια λεπτομερής περιγραφή για κάθε cluster και του πιθανοτικού προφίλ της.

4.2.6 Ανάλυση τόξων (arc analysis)

Μια άλλη ενέργεια που μας δίνει απεικόνιση της σημαντικότητας των μεταβλητών προέρχεται από την επιλογή ***arc analysis*** .Αυτή η ανάλυση δίνει έμφαση στη σημαντικότητα του ρόλου στη δομή ολόκληρου του δικτύου.Το «πάχος» του τόξου αντιστοιχεί στη δύναμη της πιθανοτική σχέση μεταξύ των μεταβλητών



Βλέπουμε ότι τα τόξα με τις σημαντικότερες εξαρτήσεις είναι όπως αναμενόταν αυτά που συνδέουν τα clusters με το ***DEVICE*** και ***QCS*** αντίστοιχα.

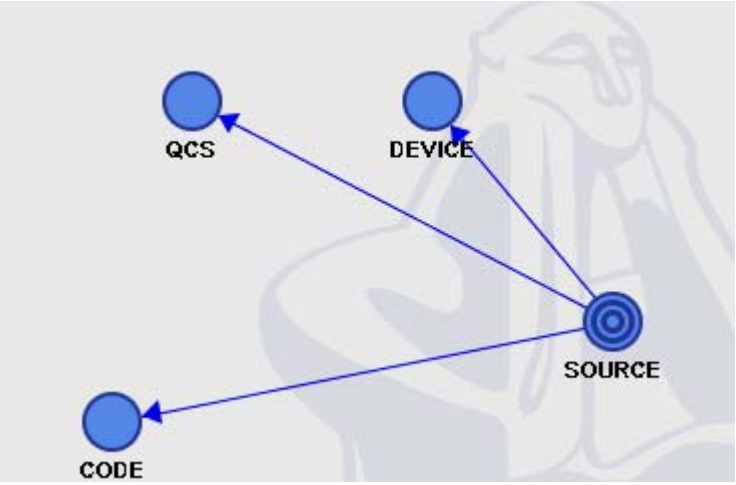
4.2.7 Επιλογή "*Supervised*" αλγόριθμου

Το Bayesialab παρέχει ένα τρίτο είδος εκμάθησης του Bayesian δικτύου ,που ανήκει στην κατηγορία των ***"supervised "*** αλγόριθμων(βασίζονται στον καθορισμό της target node πρωτού αρχίσει η όλη διαδικασία και επομένως η μορφή που θα πάρει το δίκτυο εξαρτάται κυρίως από το ποια θα είναι κάθε φορά η κεντρική μεταβλητή) .Υπάρχουν 5 αλγόριθμοι εκπαίδευσης της δομής του δικτύου και ονομαστικά αυτοί είναι :

- Naïve Bayes
- Augmented Naive Bayes
- Sons & Spouses
- Markov Blanket
- Augmented Markov Blanket

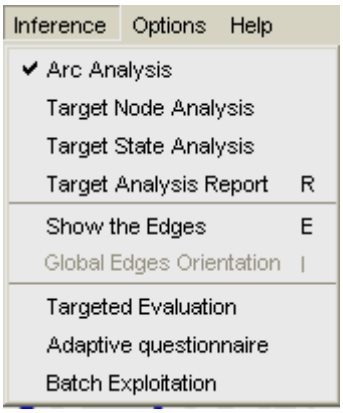
Έπειτα από δοκιμές των προαναφερθέντων μεθόδων καταλήξαμε ότι ο Augmented Naive Bayes αλγόριθμος είναι ο καταλληλότερος. Η επιλογή έγινε βάσει των δυνατοτήτων των αλγόριθμων καθώς και την ακρίβεια που υπάρχει δομή του δικτύου μια και συμπεριλαμβάνει τις σχέσεις μεταξύ των λοιπών μεταβλητών–παιδιών γνωρίζοντας τις τιμές της target node (αφού πρωτίτερα έχουμε επιλέξει ως target node την ***SOURCE***),σε αντίθεση με τον Naive Bayes αλγόριθμο.Ουσιαστικά πρόκειται για

τον Naive Bayes εμπλουτισμένο με τις πιθανοτικές εξαρτήσεις των μεταβλητών-παιδιών.Η απεικόνιση του δικτύου σύμφωνα με τον επιλεγμένο αλγόριθμο είναι η εξής:



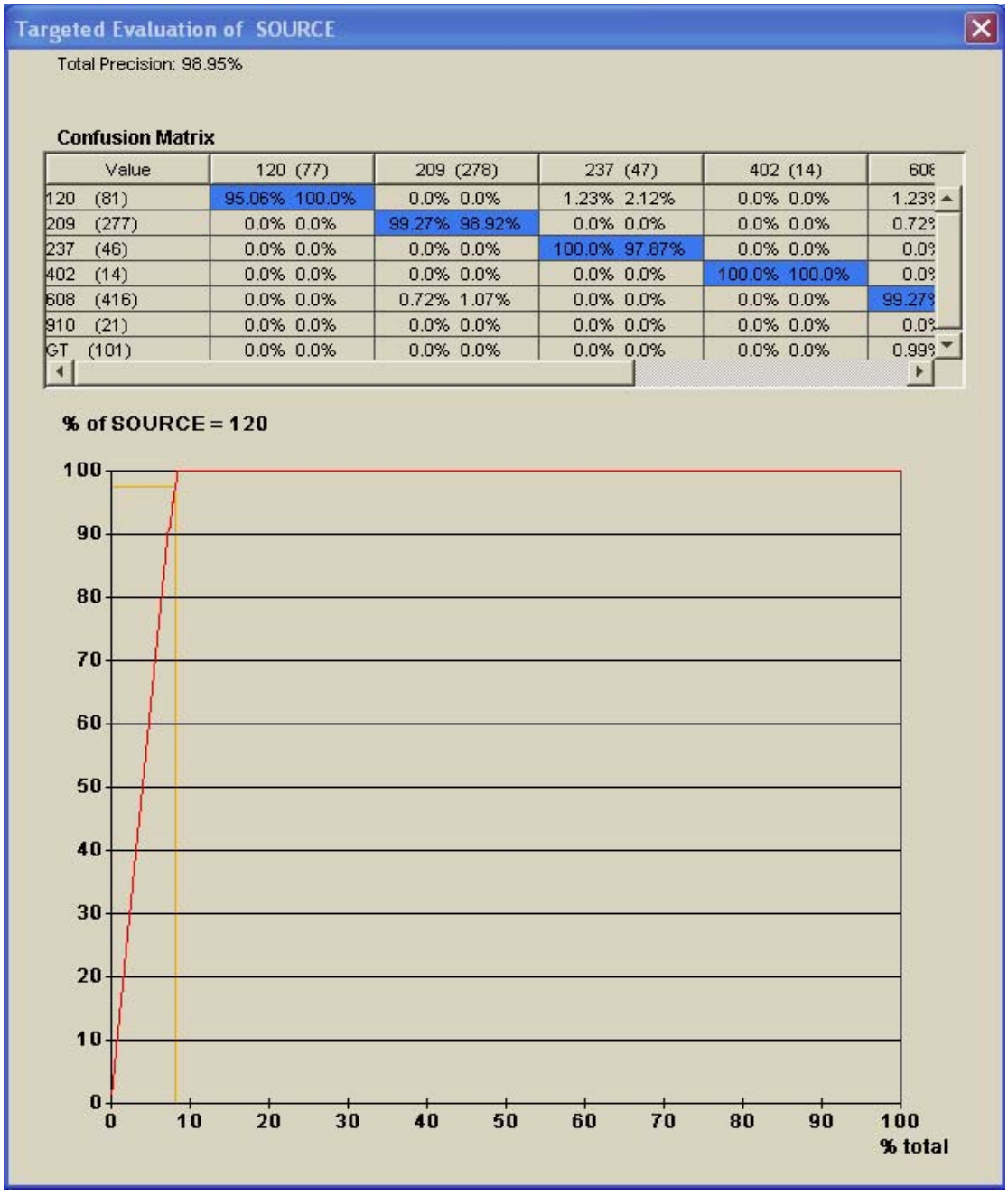
4.2.8 Εκτίμηση βάσει της κεντρικής μεταβλητής (Targeted evaluation)

Αφού πραγματοποιήθηκε η δημιουργία του μοντέλου,είμαστε σε θέση να παραθέσουμε διάφορα στοιχεία σχετικά με την απόδοση του δικτύου,μέσω του **inference menu** και της εντολής **Targeted Evaluation** .



Αυτή η εκτίμηση μας παρέχει:
- την ολική ακρίβεια του δικτύου
- τον λεγόμενο confusion matrix ο οποίος υπολογίζει τις αποδόσεις του μοντέλου για κάθε κατάσταση(όπου το πρώτο ποσοστό καταδεικνύει την αξιοπιστία του δικτύου,μέτρο των cases με σωστή πρόβλεψη ,ενώ ένα δεύτερο ποσοστό συμβολίζει την ακρίβεια ,μέτρο των cases όπου η αληθινή κατάσταση προβλέφθηκε σωτά)
- την αποκαλούμενη lift curve σχετική της καταστάσεως της target node.Αντιπροσωπεύει το ποσοστό εντοπισμού της καταστάσεως της target (Y-άξονας) σε σχέση με τον αριθμό των cases που ήδη έχουμε χρησιμοποιηθεί(X-άξονας).Τα cases ταξινομούνται ανάλογα με την ***target modality probability***(τα cases με τη μεγαλύτερη τιμή εισέρχονται πρώτα στη λίστα).Επιπλέον αν η κατάσταση της target αντιστοιχεί στο 10% των cases ,τότε η βέλτιστη lift curve έχει για τον Y-άξονα

τιμή 100%,ενώ στον X-άξονα έχει 10%(αφού όλα τα cases που αναφέρονται στην target δεν εμπεριέχουν κάποιο λάθος).Στο δικό μας δίκτυο έχουμε:

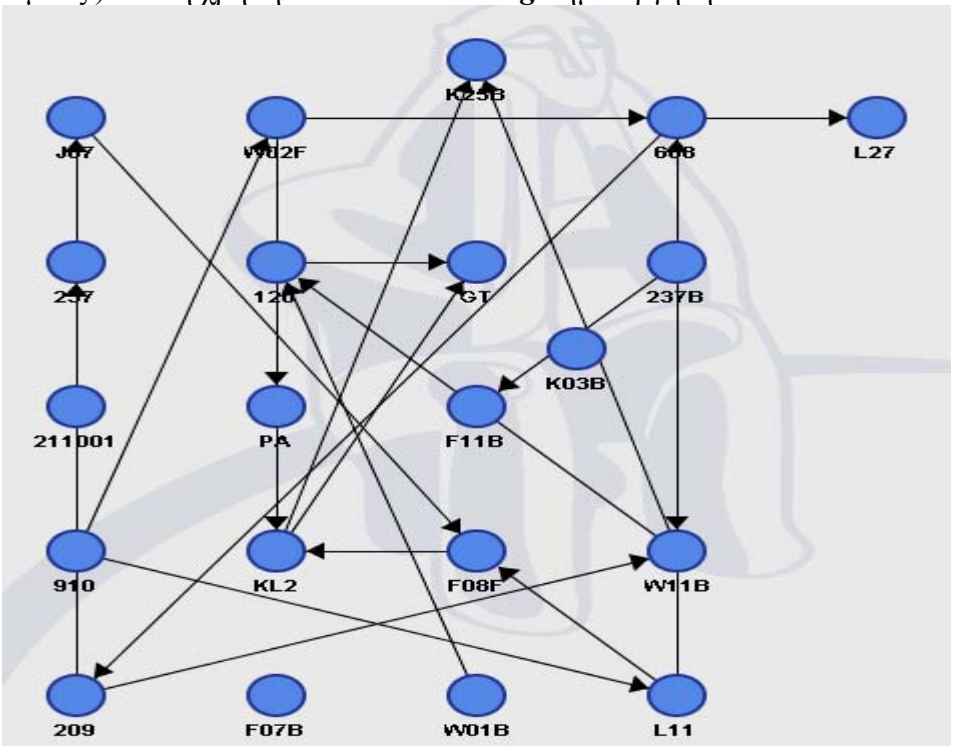


Το γεγόνος της επιλογής του Augmented Naive Bayes για την καλύτερη απεικόνιση του δικτύου ,μπορεί να επιβεβαιωθεί απλά δοκιμάζοντας τους και κατόπιν ελέγχοντας το ***Targeted Evaluation*** της.

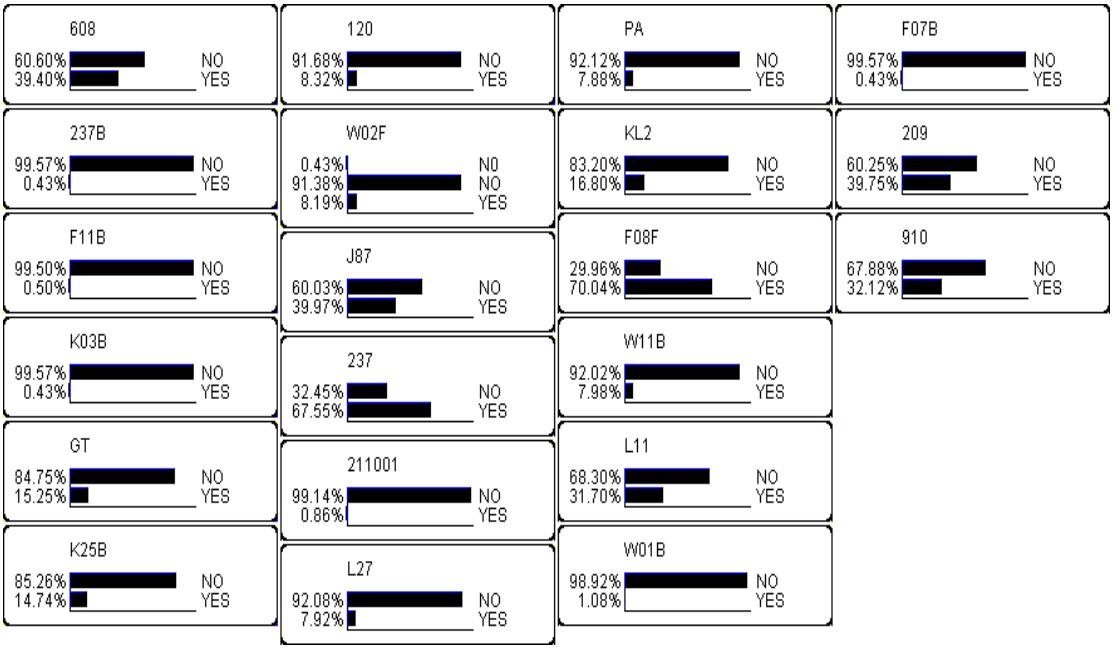
4.2.9 Εναλλακτική απεικόνιση δικτύου

Μια εναλλακτική απεικόνιση του δικτύου γίνεται κατασκευάζοντας μια βάση δεδομένων με τις 21 δημοφιλέστερες καταστάσεις των κύριων μεταβλητών και δημιουργόντας ένα δίκτυο που εκφράζει τις πιθανές εξαρτήσεις ανάμεσά τους.

Πρόκειται για δυαδική μορφή καταστάσεων (YES,NO), όπου στη τιμή YES αντιστοιχεί στην εμφάνιση της μεταβλητής.Ακολουθώντας την ίδια διαδικασία ,τόσο στην εισαγωγή των δεδομένων όσο και στην επιλογή του κατάλληλου αλγόριθμου (εδώ δεν υπάρχει κάποια target node και επομένως αποκλείουμε από τις πιθανές εκλογές τους αλγόριθμους που σχετίζονται με τους *"supervised " learning* αλγόριθμους).Με τη χρήση του ***taboo ordering*** δημιουργήθηκε το ακόλουθο δίκτυο:



Ενώ οι προβλέψεις για τις κατανομές των μεταβλητών ήταν:

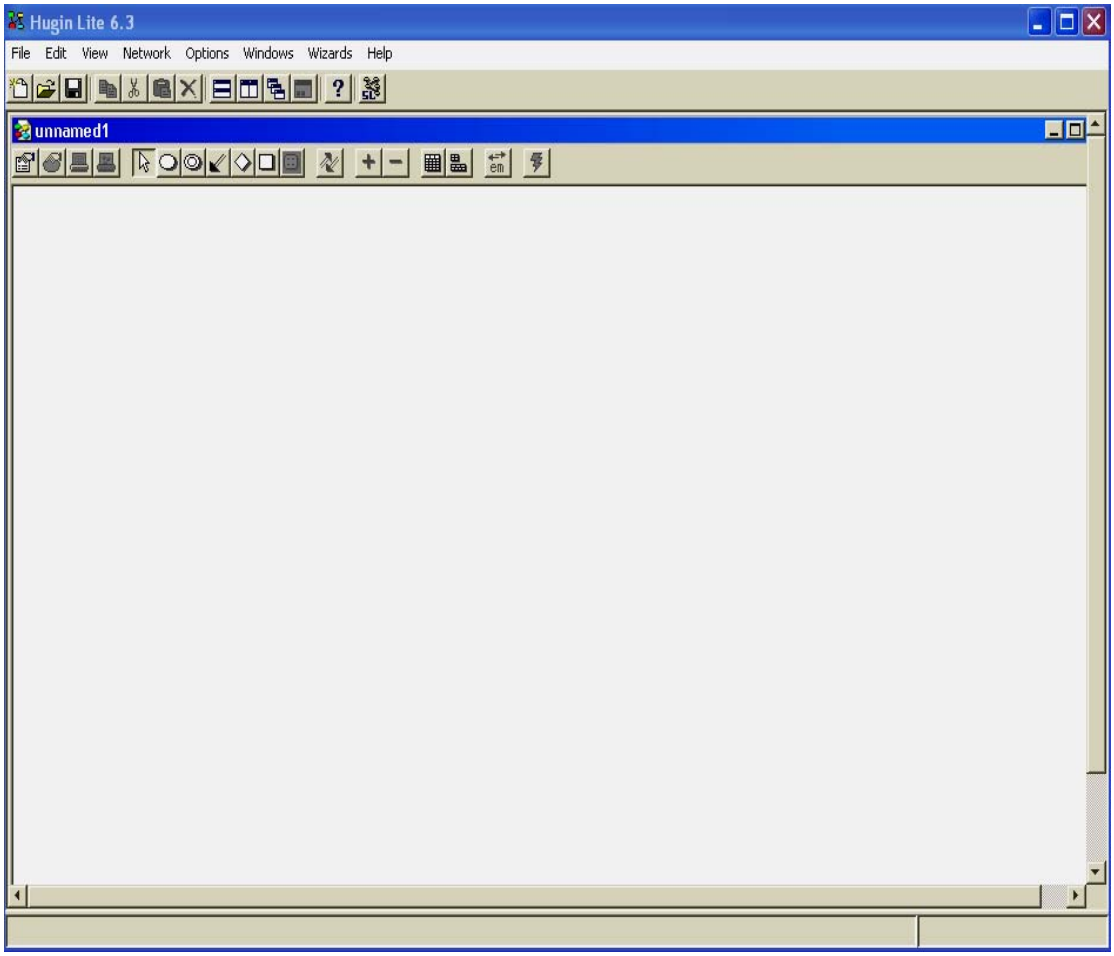


4.3 Εφαρμογή του Hugin για την κατασκευή του δικτύου

4.3.1 Εισαγωγή δεδομένων

Ξεκινάμε την κατασκευή του δικτύου έχοντας ως αφετηρία των ενεργειών μας το ***structure learning*** δηλαδή τη δημιουργία του δικτύου εισάγωντας βάσεις δεδομένων.

Αρχικά το γραφικό περιβάλλον του προγράμματος έχει την εξής μορφή:



Θα πρέπει σ'αυτό το σημείο να γίνει η υπενθύμιση του γεγονότος της απουσίας της πλήρης έκδοσης του προγράμματος που δημιουργεί αυτομάτως ένα ουσιώδη περιορισμό ·την εισαγωγή μεγάλου αριθμού μεταβλητών.Ο μέγιστος επιτρεπόμενος αριθμός είναι 25 και επομένως δεν έχουμε τη δυνατότητα της πλήρους απεικόνισης του δικτύου μια και σε ολοκληρωμένη μορφή του θα απαρτίζονταν από 129 μεταβλητές.

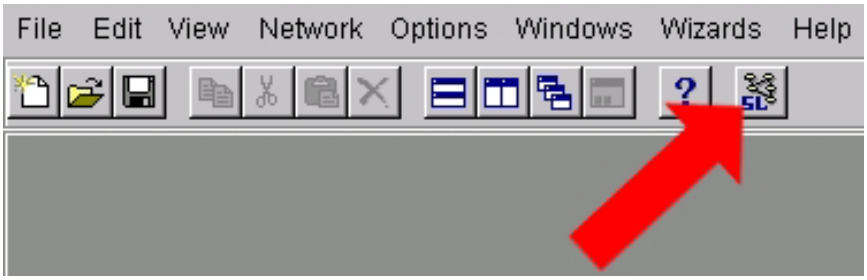
Αν'αυτού το πρώτο βήμα ήταν η δημιουργία μιας συμβατής με το πρόγραμμα βάσης δεδομένων που έχει καταγεγραμμένα γεγονότα που αναφέρονται στις 21

σημαντικότερες(πάντα σε συχνότητα εμφάνισης).Μερικές γραμμές αυτής της βάσης φαίνονται παρακάτω:

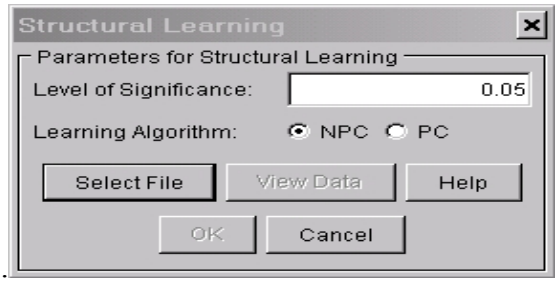
```
J87,237,211001,910,209,W02F,120,PA,KL2,F07B,K25B,GT,F11B,F08F,W01B,608,237B,K03B,W11B,L11,L27
YES,YES,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO
YES,YES,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO
YES,YES,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO
YES,YES,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO
YES,YES,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO
YES,YES,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO
YES,YES,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO
YES,YES,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO
YES,YES,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO
YES,YES,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO
YES,YES,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO
YES,YES,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO
YES,YES,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO
YES,YES,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO
YES,YES,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO
YES,YES,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO
YES,YES,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO
YES,YES,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO
YES,YES,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO
YES,YES,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO
YES,YES,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO
YES,YES,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO
YES,YES,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO
YES,YES,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO
```

όπου στην πρώτη γραμμή αναφέρονται οι μεταβλητές.

Η επιλογή του ***structural learning*** είναι διαθέσιμη είτε στο μενού ***"File"*** ,όπως επίσης υπάρχει και πάνω στο ***Tool bar*** του ***Main Window***

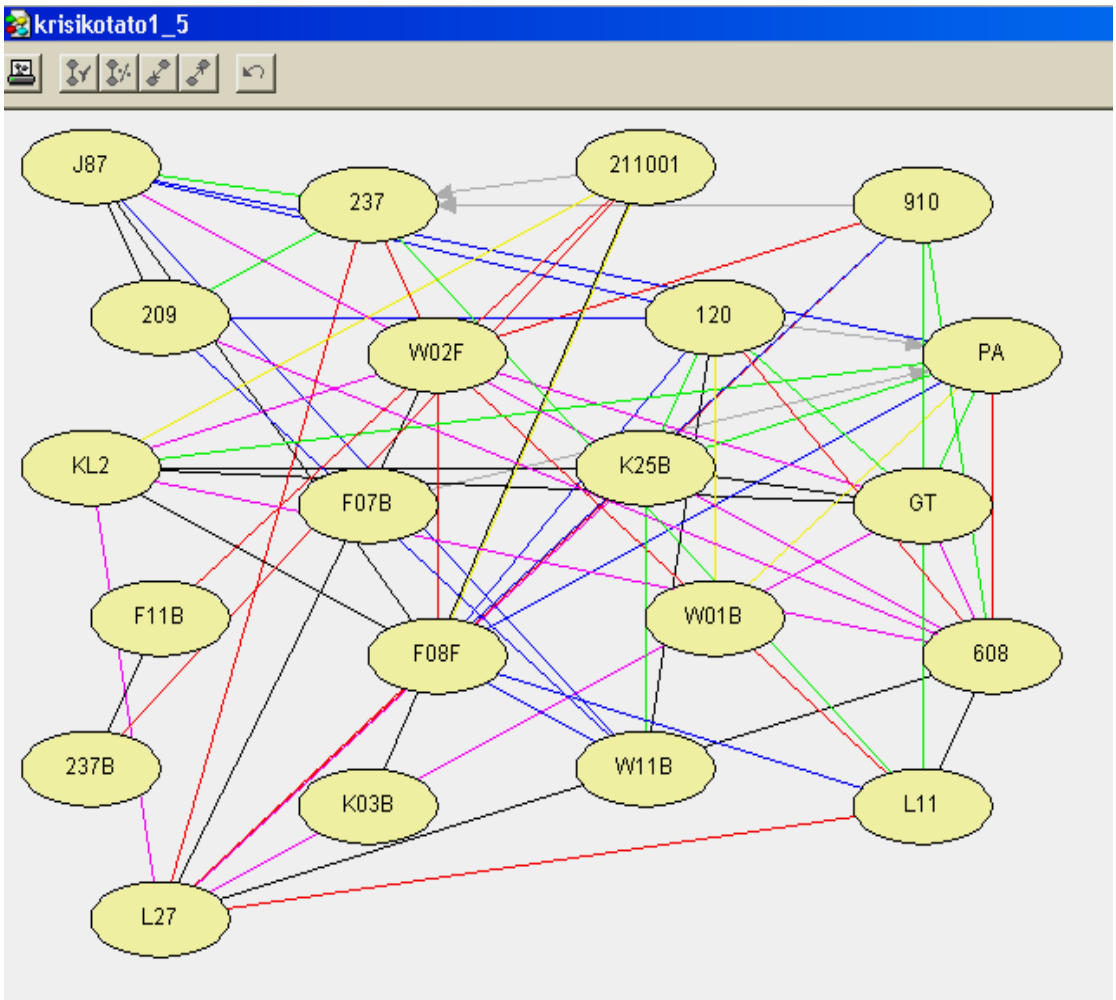


Αφού πατηθεί το κουμπί εμφανίζεται το ***"structure learning dialog"*** .Σ'αυτό σημείο ζητείται από το πρόγραμμα να οριστεί το ***"Level of Significance***" το οποίο προσδιορίζει το επίπεδο σημαντικότητας των στατιστικώς ανεξαρτήτων tests που διενεργούνται κατά τη διάρκεια του ***structural learning*** και παίρνει τιμές στο διάστημα : (0,1).Στη συνέχεια επιλεγουμε το αρχείο που περιέχει τη βάση και τέλος αν και όχι λιγότερο σημαντικό θα πρέπει να επιλέξουμε τοναλγόριθμο με τον οποίο θα γίνει η «εκμάθηση».Θα αναφερθούμε εκτενώς στο τελευταίο μέρος του παρόντος κεφαλαίου.Η επιλογή μας θα είναι ο αλγόριθμος ***NPC(***necessary path condition).Κλείνουμε αυτό το στάδιο πατώντας OK

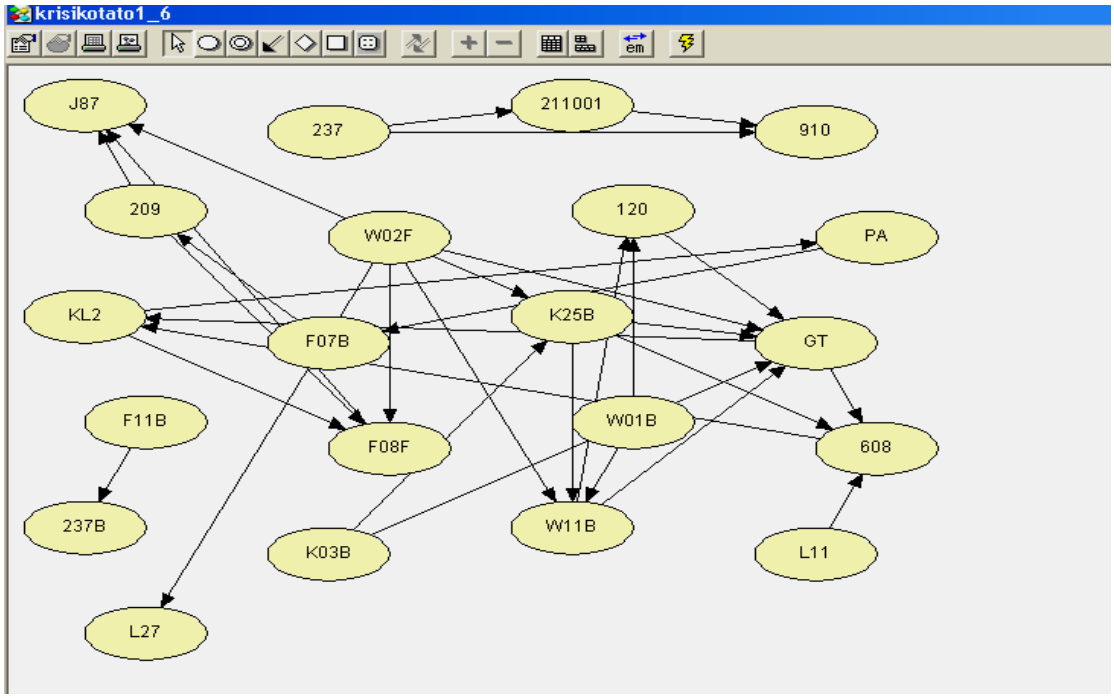


4.3.2 Σχηματοποίηση δικτύου

Μετά την επιλογή του ***NPC*** μες στο δίκτυο θα υπάρχουν αβέβαιοι σύνδεσμοι ή σύνδεσμοι που η κατεύθυνσή τους δεν είναι εφικτό να καθοριστεί με βεβαιότητα και εδώ ο χρήστης με την βοήθεια ένος γραφικού διαισθητικού μηχανισμού καλείταινα επιλύσει αυτές τις αβεβαιότητες.Η επόμενη εικόνα παρουσιάζει το αρχικό αποτέλεσμα που είχε η χρησιμοποίηση του ***NPC*** στη δική μας περίπτωση.



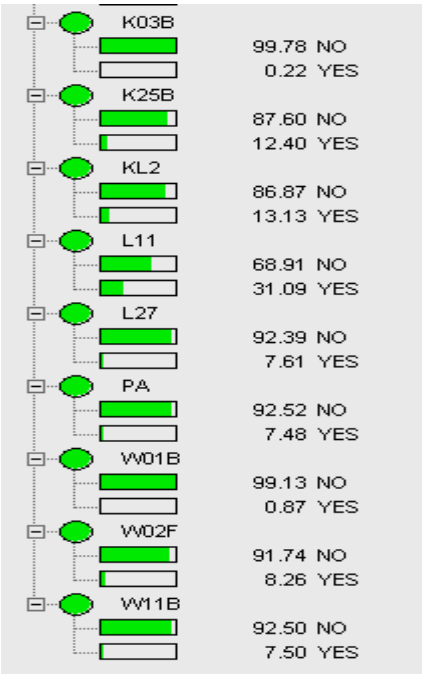
Η τελική μορφή του δικτύου ποικίλει ανάλογα με τις επιλογές που θα κάνει ο χρήστης.Η μορφή που επιλέξαμε είναι η παρακάτω:

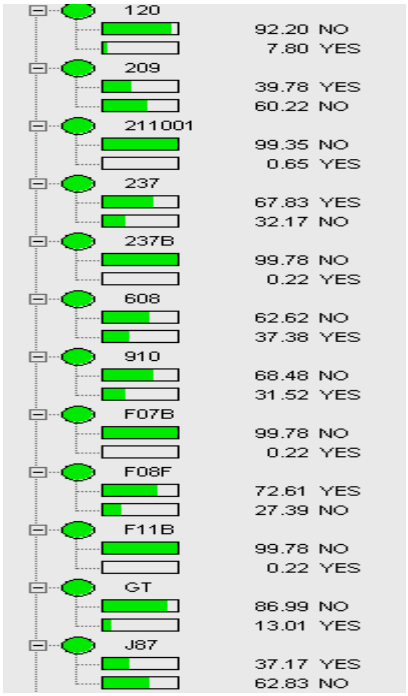


4.3.3 Εύρεση των κατανομών πιθανότητας των μεταβλητών

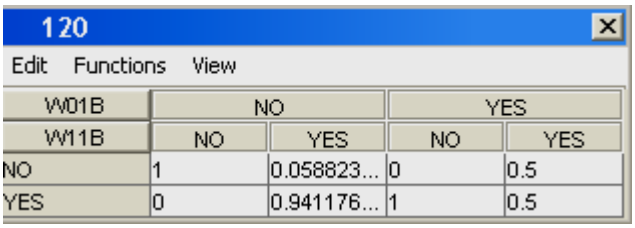
Όταν η δομή του δικτύου δημιουργηθεί ,τότε οι κατά συνθήκη κατανομές πιθανότητας των μεταβλητών μπορούν να υπολογιστούν από τα διαθέσιμα δεδομένα κάνοντας κάνοντας χρήση του ***EM-learning*** αλγόριθμου. Η επιλογή αυτή υπάρχει στο μενού του ***NETWORK*** καθώς και στο ***edit mode*** του ***tool bar*** του ***Main Window(***  ***)***

Εκτελώντας ανάλογη διαδικασία με προηγουμένως(δίνοντας το αρχείο με τα δεδομένα καθως και τον επιθυμητό αριθμό επαναλήψεων) υπολογίσαμε τις κατανομές των μεταβλητών.





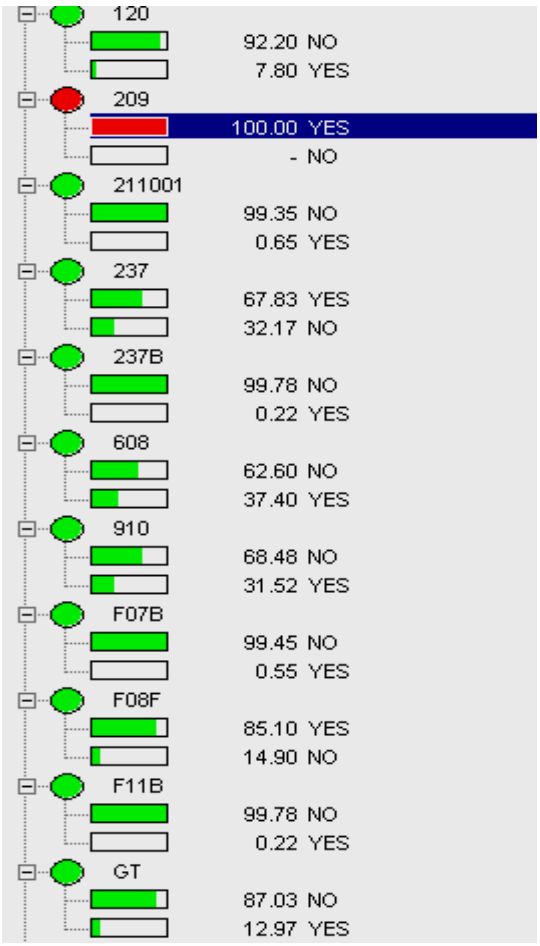
Επίσης μέσω του κουμπιού  που υπάρχει στο ***edit mode*** μπορούμε να δούμε τους κατα συνθήκη πιθανότητες πίνακες (***CPT***). Έστω ότι θέλουμε να δούμε τον πίνακα της μεταβλητής ***120.***Αύτος έχει τη μορφή:

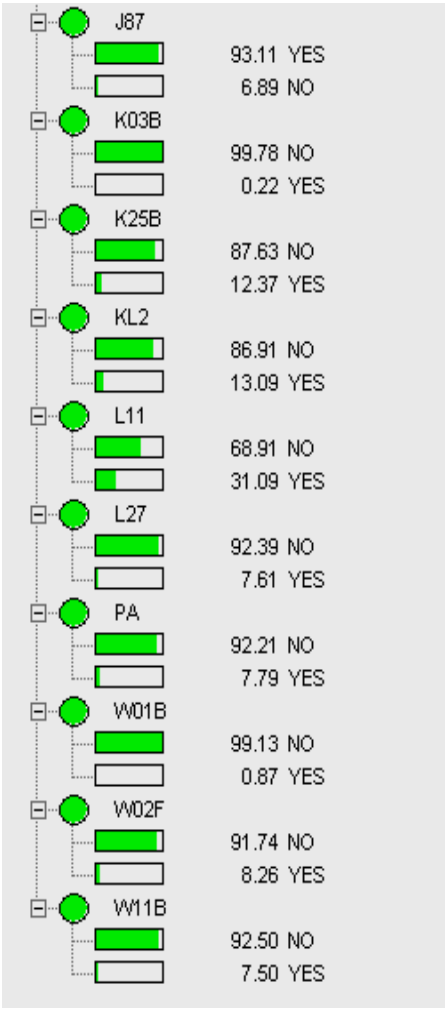


μια και η μεταβλητή 120 είναι εξαρτημένη από τις μεταβλητές-γονείς ***WO1B*** και ***W11B.***

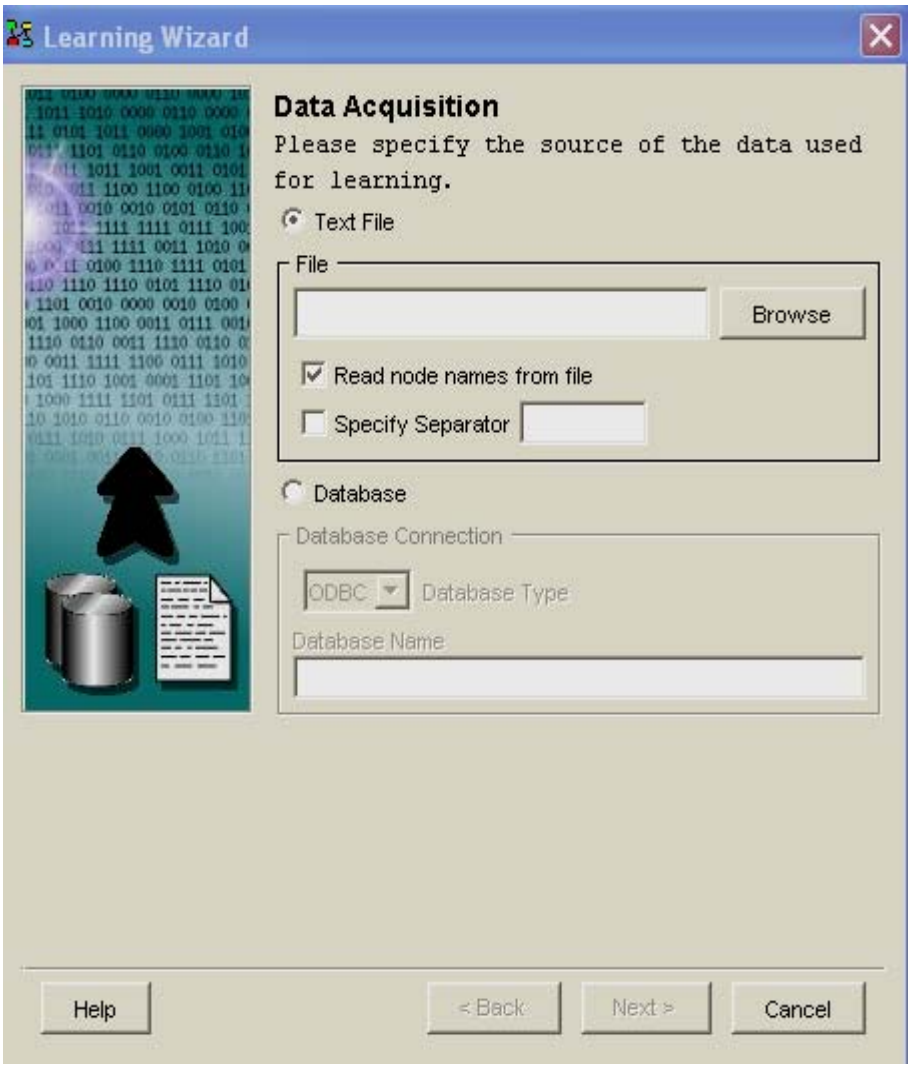
4.3.3 Η διαδικασία ανανέωσης (Propagation algorithm)

Επίσης υπάρχει η διαδικασία της ανανέωσης της κατανομής πιθανότητας σε περίπτωση νέων παρατηρήσεων.Για παράδειγμα θεωρώντας ότι η κατάσταση της μεταβλητής **209** είναι γνωστή , τότε οι νέες κατανομές γίνονται:



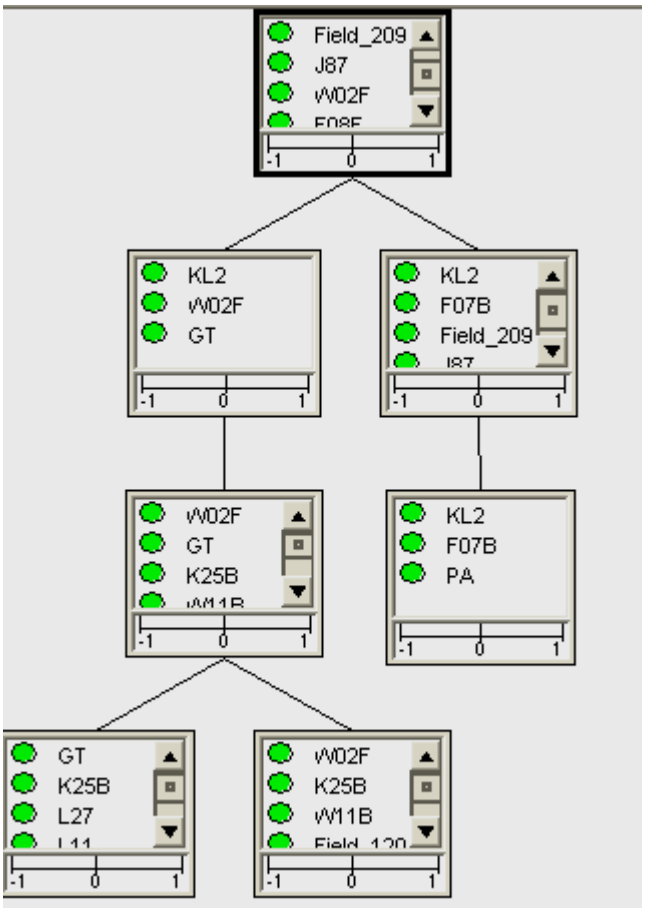


Βέβαια ένας άλλος τρόπος προσδιορισμού τόσο των κατανομών όσο και της δομής του δικτύου,(ακόμα και στην περίπτωση που δεν είναι διαθέσιμη η δομή του δικτύου) μπορούμε να κάνουμε χρήση του Learning Wizard :



4.3.5 Junction tree

Επίσης το πρόγραμμα παρέχει και την απεικόνιση του ***junction tree*** ,η οποία επιλογή ενεργοποιείται όταν είμαστε σε ***run mode***



Για τέλος του κεφαλαίου παρουσιάζονται μερικές λεπτομέρειες για τους αλγόριθμους που χρησιμοποιούνται στο ***Hugin***

4.4 Πληροφορίες για τους NPC,PC αλγόριθμους

Δύο αλγόριθμοι είναι διαθέσιμοι για την εκμάθηση δομών: Ο αλγόριθμος **PC** και ο αλγόριθμος**NPC**. **Ο *NPC*** αντιπροσωπεύει *τον* **Necessary path condition** και είναι ένα κριτήριο που αναπτύχθηκε από τους ερευνητές της Siemens στο Μόναχο για την επίλυση μερικών προβλημάτων που αντιμετωπίζουν οι αλγόριθμοι μάθησης βασισμένοι σε περιορισμούς(constraint-based learning algorithm) όπως τον αλγόριθμο ***PC***. Τα βασικά μηχανήματα είναι τα ίδια στο ***PC*** και τους αλγορίθμους ***NPC*** (δηλ., είναι και οι δύο βασισμένοι στην παραγωγή ενός ‘’σκελετού’’ που παράγεται μέσω των στατιστικών δοκιμών για την υπό όρους ανεξαρτησία).

Ο αλγόριθμος ***NPC*** επιδιώκει να επισκευάσει τις ανεπάρκειες του αλγορίθμου ***PC***, οι οποίες εμφανίζονται ειδικά στα περιορισμένα σύνολα στοιχείων. Η λύση που παρέχεται από τον αλγόριθμο ***NPC*** είναι βασισμένη στο συνυπολογισμό ενός κριτηρίου γνωστού ως **necessary path condition** . Αυτό το κριτήριο αποτελεί τη βάση για την έννοια των ***Ambiguous regions*** (διφορούμενων περιοχών) , οι οποίες παρέχουν με τη σειρά τους , μια ‘γλώσσα’’ για την επιλογή μεταξύ των συνόλων αλληλοεξαρτώμενων αβέβαιων συνδέσεων. Η επίλυση των ambiguous regions εκτελείται στην αλληλεπίδραση με το χρήστη.

Στους ***constraint-based learning algorithms***, ο σκελετός της γραφικής παράστασης κατασκευάζεται από τη μη συμμετοχή ενός συνδέσμου στο παραγόμενου δικτύου όποτε οι ανταποκρινόμενοι κόμβοι βρίσκονται να είναι υπό συνθήκη ανεξάρτητοι.. Μπορούν, εντούτοις, να υπάρξουν ασυνέπειες μεταξύ του συνόλου των υπό όρους καταστάσεων ανεξαρτησίας και εξάρτησης (***CIDs***) που προέρχονται από τα περιορισμένα σύνολα στοιχείων. Δηλαδή δεν μπορεί να απεικονιστούν ταυτόχρονα όλα τα CIDs. Οι ασυνέπειες υποτίθεται ότι προήλθαν απλώς από το θόρυβο δειγματοληψίας(sampling noise) π.χ., ακόμα υποθέτουμε ότι υπάρχει ένας τέλειος χάρτης ανεξαρτησίας(***perfect independence map***) της (άγνωστης) κατανομής πιθανότητας από την οποία τα δεδομένα παράγονται.

Ο αριθμός ασυνεπειών στο σύνολο των ***CIDs*** απεικονίζει τη δομική αβεβαιότητα του μοντέλου(***structural model uncertainty***). Κατά συνέπεια, ο αριθμός αβεβαιοτήτων είναι ένα μέτρο εμπιστοσύνης για τη εκπαιδευμένη δομή και μπορεί υπό αυτήν τη μορφή να χρησιμοποιηθεί ως ένδειξη για το κατά πόσο χρησιμοποιήθηκαν τα κατάλληλα δεδομένα για να την εκμάθηση. Τα ασυμβίβαστα CIDs παράγουν τις πολλαπλάσιες λύσεις κατά την πρόκληση μιας προσανατολισμένης ακυκλικής γραφικής παράστασης (DAG) από τους. Αυτές οι λύσεις διαφέρουν όσον αφορά το σύνολο συμπεριλαμβανόμενων συνδέσεων.

Για να επιλύσει τις ασυνέπειες, ο αλγόριθμος ***NPC*** στηρίζεται στην αλληλεπίδραση του χρήστη, όπου ο χρήστης έχει την ευκαιρία να αποφασίσει σχετικά με την κατεύθυνση των undirected συνδέσεων και να επιλύσει τις ***ambiguous regions(ασαφής περιοχές)***

4.4.1 Necessary path condition

Ανεπίσημα, ο **Necessary path condition** λέει ότι δύο μεταβλητές ***X*** και ***Y*** είναι ανεξάρτητες υπό όρους στο καθορισμένο σύνολο ***S***, χωρίς το κατάλληλο υποσύνολο του ***S*** για το οποίο αυτό ισχύει, πρέπει να υπάρξει μια πορεία μεταξύ του ***X*** και κάθε ***Z*** στο ***S*** (χωρίς να διασχίζει το ***Y***) και μεταξύ του ***Y*** και κάθε ***Z*** στο ***S*** (που δεν διασχίζει ***X***). Διαφορετικά, ο συνυπολογισμός του ***Z*** στο ***S*** είναι ανεξήγητος. Κατά συνέπεια, για να ισχύει μια κατάσταση ανεξαρτησίας, διάφορες συνδέσεις πρέπει για να είναι παρούσες στη γραφική παράσταση.

4.4.2 Ambiguous regions

Όταν η απουσία μιας σύνδεσης, ***a,*** εξαρτάται από την παρουσία μιας άλλης σύνδεσης, ***b***, και αντίστροφα, ορίζουμε το ***a*** και το ***b*** ότι είναι αλληλοεξαρτώμενοι. Και το ***a*** και το ***b*** αποτελούν τι αποκαλούμενες *αβέβαιες συνδέσεις(**uncertain links**)*. Μια *ambiguous region* είναι ένα μέγιστο σύνολο αλληλοεξαρτώμενων συνδέσεων(***inter-dependent links***). Δηλαδή μια ambiguous region αποτελείται από ένα σύνολο αβέβαιων συνδέσεων.

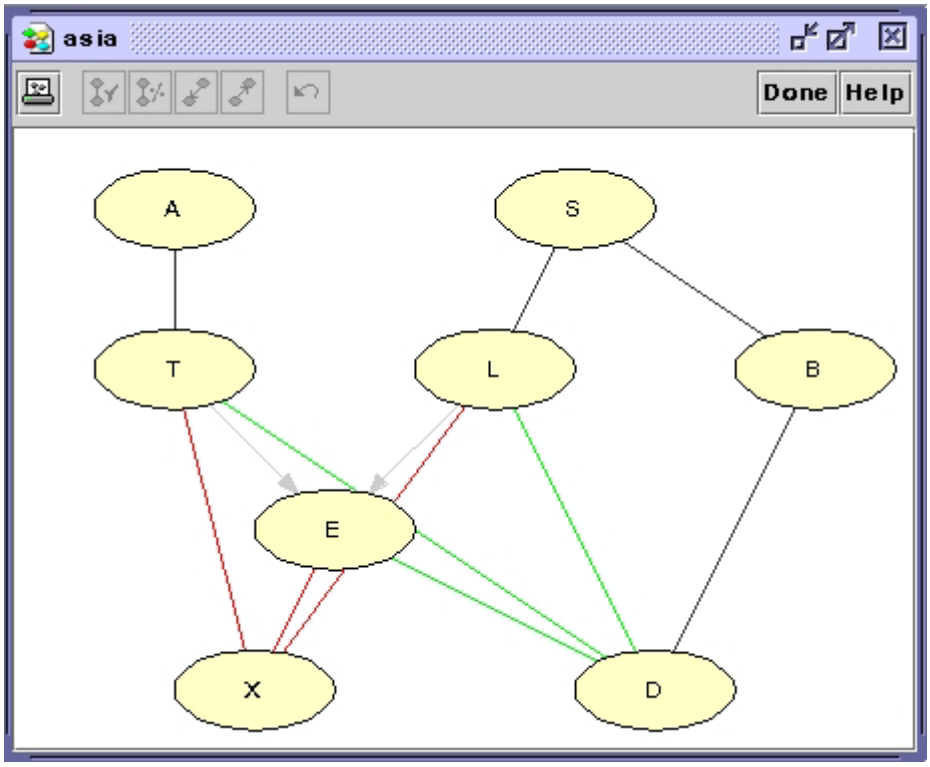
Ο κύριος στόχος είναι να ληφθούν όσο το δυνατόν λίγες και μικρότερες ambiguous regions. Πρέπει να σημειωθεί ότι οι καθορισμένες σχέσεις μεταξύ των μεταβλητών θα παραγάγουν επίσης ambiguous regions

Εάν υπάρχουν μερικές αβέβαιες συνδέσεις (ή συνδέσεις που πρέπει να προσανατολιστούν), παρέχεται στο χρήστη η δυνατότητα τις πληροφορίες ως προς

τον τρόπο με τον οποίο οι ambiguous regions πρέπει να επιλυθούν.Ένα ειδικό, διαισθητικό γραφικό περιβάλλον παρέχεται για αυτήν την αλληλεπίδραση.

4.4.3 Επίλυση των ambiguous regions και της έλλειψης προσανατολισμού

Κατά χρησιμοποίηση του αλγορίθμου ***NPC***, ένα διαισθητικό γραφικό περιβάλλον παρέχεται για την επίλυση των δομικών αβεβαιοτήτων που βρίσκονται (ενδεχομένως) από τον αλγόριθμο. Το σχήμα 1 παρουσιάζει ένα παράδειγμα μιας δομής με τις εκκρεμείς αβεβαιότητες. και τις δυνατότητες που παρέχει το Hugin



Σχήμα 1: Οι δομικές αβεβαιότητες που βρίσκονται από τον αλγόριθμο ***NPC***. Οι μαύρες (undirected) συνδέσεις είναι επιλέξιμες και ο προσανατολισμός τους μπορεί να καθοριστεί από το χρήστη. Οι συνδέσεις κάθε ambiguous regions σχεδιάζονται με το ίδιο χρώμα . Αυτές οι συνδέσεις είναι επίσης επιλέξιμες και μπορούν να αφαιρεθούν ή να κρατηθούν, ανάλογα με τη ενέργεια που εκτελείται από το χρήστη.

Εάν δεν είναι επιθυμητή η παροχή τέτοιων πληροφοριών , απλά πατάμε το κουμπί ***Done*** και ο αλγόριθμος ***NPC*** θα επιλύσει τις αβεβαιότητες. Σημειώστε, εντούτοις, ότι ο προσανατολισμός για τις undirected συνδέσεις θα αποφασιστεί σε τυχαία βάση, και ότι για εκείνες τις ambiguous regions χωρίς παρεχόμενες πληροφορίες, όλες οι αβέβαιες συνδέσεις θα αφαιρεθούν, ενδεχομένως με συνέπεια ένα ‘’φτωχό’’ μοντέλο.

4.4.4 Επίλυση απροσανατολισμένων συνδέσμων

Αντί να καθορίζεται τυχαία ο προσανατολισμός των συνδέσεων της εκπαιδευμένης δομής που δεν μπορούν να καθοριστούν αυτόματα από τα δεδομένα, ο αλγόριθμος ***NPC*** δίνει στο χρήστη την ευκαιρία να καθορίσει τον προσανατολισμό τέτοιων συνδέσεων.

Οι undirected links (που δεν ανήκουν στις ambiguous regions) σχζεδιάζονται με μαύρο χρώμα . Όταν επιλέγεται, μια τέτοια σύνδεση τονίζεται και δύο κουμπιά προσανατολισμού ενεργοποιούνται:  ή  ή 

Σχήμα 2: Οι διαφορετικές εμφανίσεις των κουμπιών κατευθυντικότητας

Το ποιά από τις ανωτέρω εμφανίσεις του ζευγαριού των κουμπιών χρησιμοποιηθεί εξαρτώνται από τη σχετική θέση των κόμβων της επιλεγμένης σύνδεσης. Κατά συνέπεια, πρέπει να είναι εύκολα προφανές ποιο από τα δύο κουμπιά πρέπει να επιλεχτεί για την επίτευξη του επιθυμητού προσανατολισμού για την επελεγμένη σύνδεση.

Σημειώστε ότι η ανάθεση του προσανατολισμού σε μια σύνδεση μπορεί να αναγκάσει άλλες συνδέσεις να προσανατολιστούν αυτόματα. Εάν, παραδείγματος χάριν, υπάρχουν undirected links μεταξύ των μεταβλητών ***X*** και ***Y***, μεταξύ του ***Y*** και του ***Z,*** και μεταξύ του ***X*** και του ***Z***, κατόπιν επιβάλλοντας τις ακόλουθες κατευθύνσεις $X \rightarrow Y$ και $Y \rightarrow Z$ τότε συνεπάγεται ότι $X \rightarrow Z$ διαφορετικά, ένας προσανατολισμένος κύκλος θα εμφανιζόταν.

4.4.5 Επίλυση των (ambiguous regions)διφορούμενων περιοχών

Μια ambiguous region αποτελείται από ένα σύνολο αλληλοεξαρτώμενων αβέβαιων συνδέσεων: Απουσία μιας σύνδεσης σε μια τέτοια περιοχή εξαρτάται από την παρουσία ενός ή και περισσοτέρων συνδέσμων της περιοχής και αντίστροφα.

Κάθε ambigugous region προσδιορίζεται εύκολα είναι ένα σύνολο αποτελούμενο από συνδέσεις με ίδιο χρώμα. (Σημειώστε ότι μπορούν να υπάρξουν τόσες πολλές ambiguous regions , που καθιστά δύσκολη τη διάκρσή τους μόνο από το χρωματισμό.) Επίσης, κατά επιλογή μιας σύνδεσης από μια ambigugous region όλες τη συνδέσεις της περιοχής θα τονιστούν .
Όταν μια σύνδεση μιας ambigugous region επιλεχτεί, τα κουμπιά της συνυπολογισμού/αποκλισμού ενεργοποιούνται: 

Όταν μια απόφαση ληφθεί σχετικά με το αν μια σύνδεση πρέπει να είναι παρούσα ή απούσα, κάθε μια από τις άλλες συνδέσεις της ambigugous region θα επηρεαστεί με έναν από παρακάτω τρόπους:

- Παραμένει ανεπιρρέαστος.
- Εξαφανίζεται
- Μετατρέπεται σε undirected σύνδεση, που δεν ανήκει στην περιοχή πλέον.
- Μετατρέπεται σε προσανατολισμένη σύνδεση, που δεν ανήκει στην περιοχή άλλο.

Ποια από αυτές τις συνέπειες θα παρατηρηθεί εξαρτάται από τις υπό όρους καταστάσεις ανεξαρτησίας και εξάρτησης ***(CIDs)*** που βρίσκονται από τις στατιστικές δοκιμές που εκτελούνται από τον αλγόριθμο ***NPC***.

4.4.6 Αλγόριθμος *PC*

Ο αλγόριθμος ***PC*** είναι ένας αλγόριθμος που μπορεί να μειώσει αυτήν την πολυπλοκότητα παρά πολύ. Είναι μια από τις δημοφιλέστερες προσεγγίσεις βασισμένες σε περιορισμούς.

Βασική υπόθεση

- Οι σχέσεις ανεξαρτησίας έχουν μια τέλεια αντιπροσώπευση από ένα ***DAG***
- Έχουμε στατιστικές δοκιμές δεν έχουν κανένα λάθος
- Διαθέτουμε πολύ μεγάλες βάσεων δεδομένων

Υπό αυτούς τους όρους, ο αλγόριθμος θα ανακαλύψει και ισοδύναμο Μπεϋζιανό δίκτυο.

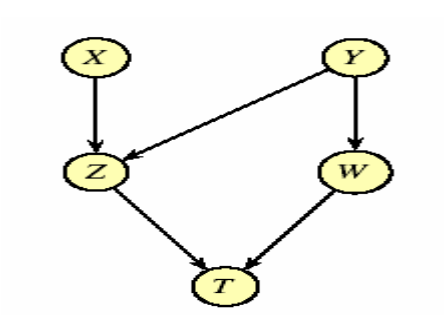
Στατιστικές δοκιμές

Ο αλγόριθμος είναι βασισμένος στην αναζήτηση των αληθινών σχέσεων ανεξαρτησίας της μορφής: $I(Xi;XjjA)$ όπου *το* ***A*** είναι ένα υποσύνολο των μεταβλητών Μπορεί να λειτουργήσει με οποιαδήποτε πηγή που παρέχει αυτό το είδος πληροφοριών.

Η δομή του αλγόριθμου

- Βρείτε ένα γραφικό σχέδιο(patern) (gp): an undirected graph
- Βρείτε μερικές ***head to head*** συνδέσεις δοκιμάζοντας τις ανεξαρτησίες
- Προσανατολίστε τις υπόλοιπες συνδέσεις χωρίς την αναπαραγωγή κύκλων

Graph pattern:The basic condition



Δύο κόμβοι,X και Y ,συνδέονται αν και μόνο αν δεν υπάρχει υποσύνολο S_{XY} από το σύνολο των τόξων τέτοιο ώστε $I(Xi;XjjA)$

Θα μπορούσαμε να προσπαθήσουμε να ανακαλύψουμε το σχέδιο γραφικών παραστάσεων μετά από αυτό το κριτήριο, αλλά θα είναι ανεπαρκές (πάρα πολλές δοκιμές) και ανακριβές (ρυθμίζοντας σε πολλές μεταβλητές)

4.4.7 Βρίσκοντας το Graph Pattern

Εστω V είναι το σύνολο των κόμβων και κάθε σχέση ανεξαρτησίας μπορεί να δοκκιμαστεί.Κάθε κόμβος έχει ένα σύνολο γειτονικών ADJX .

1. Start with a complete undirected graph gp

2. i = 0

3. Repeat

4. For each X $\in$ **V**

5. For each Y $\in$ **ADJX**

6. Determine if there is $S \subseteq ADJ_X - \{Y\}$ with $|S| = i$

and I(X,Y /S)

7. If this set exists

8. Make S XY = S

9. Remove X _Y link from gp

10. i = i+1

11. Until $|ADJ_X| \leq i \forall$. **X**

Κεφάλαιο 5ο

5.1 Επίλογος και κριτική αποτελεσμάτων

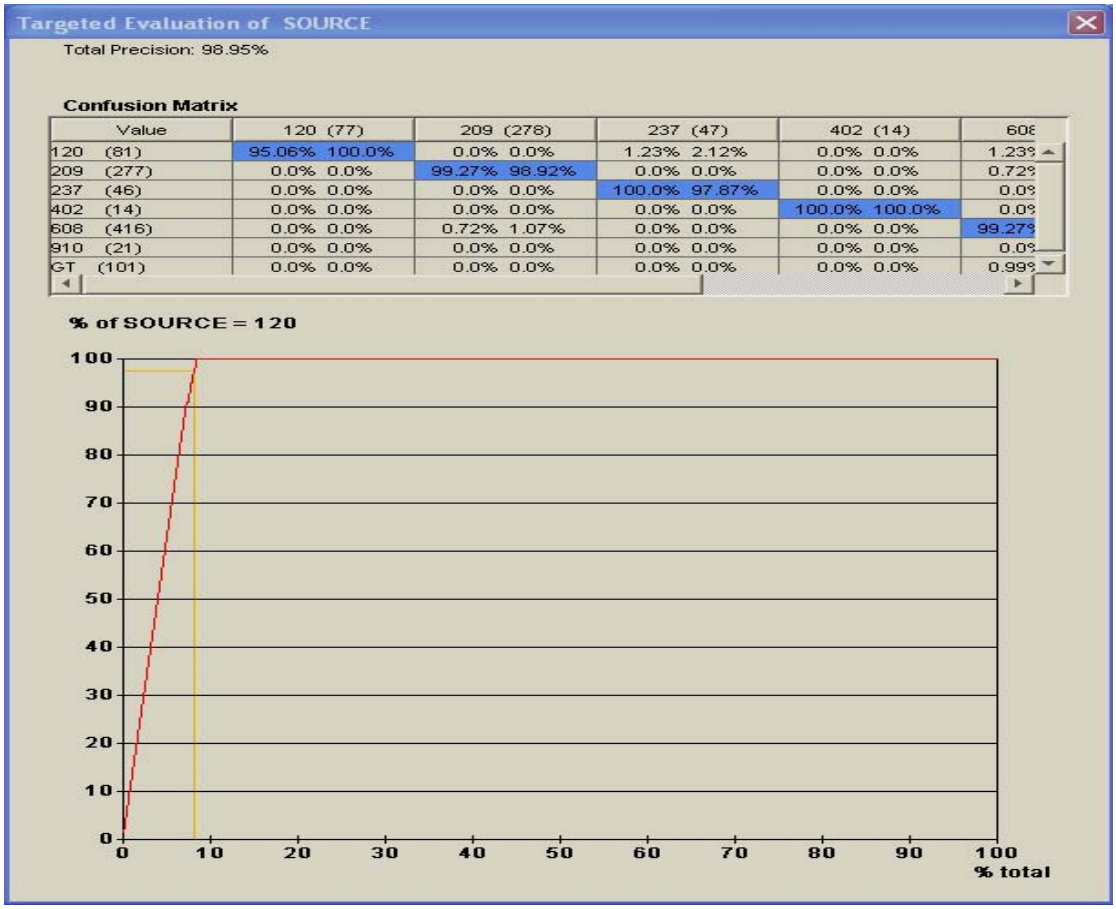
Οι υπεύθυνοι της γραμμής παραγωγής του προιόντος βρίσκονται αντιμέτωποι με το δυσεπίλυτο πρόβλημα της μεγάλης σε συχνότητα εμφάνισης μη λειτουργικών προϊόντων-πράγμα που σημαίνει την δυσλειτουργία της.

Ο στόχος μας ήταν μέσω της χρήσης της θεωρίας των Bayesian Networks ,αλλά και των δύο εκ των δημοφιλέστερων και έγκριτων λογισμικών (Hugin,Bayesialab) σ'αυτόν τον χώρο να κάνουμε κάποια πρόβλεψη για την μελλοντική συμπεριφορά των παραμέτρων που επιρεάζουν το δίκτυο και επομένως ένα αποτελεσματικό τρόπο αντιμετώπισης ενδεχόμενων αντίστοιχων προβλημάτων καθώς και βελτιωτικών κινήσεων σε τομείς όπου είναι ανεπαρκείς η παραγωγική διαδικασία(π.χ περαιτέρω εκπαίδευση των εργαζόμενων,σχολαστική και συχνότερη επισκευή των μηχανών,αλλαγή του εξωτερικού προμηθευτή κ.α).

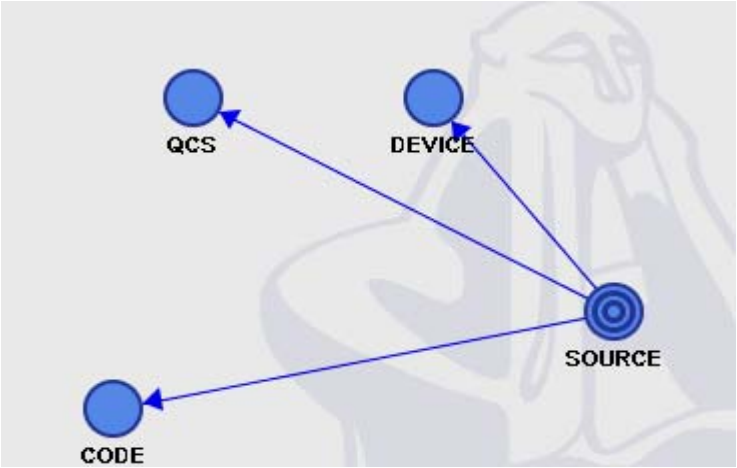
Ένα πρώτο βήμα προς αυτή την κατεύθυνση είναι ο καθορισμός της λεγόμενης ***target node*** ,στην οποία ουσιαστικά δίνουμε μεγάλη βαρύτητα,λόγω της σημαντικότητάς στη διαμόρφωση της συμπεριφοράς ολόκληρου του συστήματος.Η επιλογή μας ήταν η μεταβλητή ***Source*** μια και αυτή αναφέρεται στη πηγή που προκάλεσε το σφάλμα.

Επομένως θα ασχοληθούμε κυρίως με μία από τις κατηγορίες αλγορίθμων που υπάρχουν στο Bayesialab και πιο συγκεκριμένα ,με την κατηγορία των ***" supervised learning for the characterization of a particular variable"***.Όπως έχουμε ήδη αναφερθεί η καλύτερη προσέγγιση βάσει των ***"targeted evaluation "*** που προκύπτει έπειτα από κάθε δοκιμή των διαθέσιμων αλγόριθμων είναι ο ***"Augmented Naïve Bayes"*** μια και δίνει το μεγαλύτερο ποσοστό συνολικής ακρίβειας (98,95%),όπως και βάσει των δυνατοτήτων του αλγόριθμου καθώς και την ακρίβεια που υπάρχει δομή του δικτύου μιας και συμπεριλαμβάνει τις σχέσεις μεταξύ των λοιπών μεταβλητών–παιδιών γνωρίζοντας τις τιμές της ***target node***

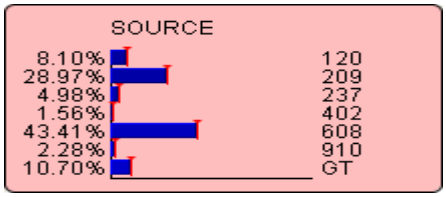
.

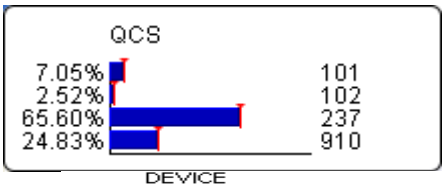


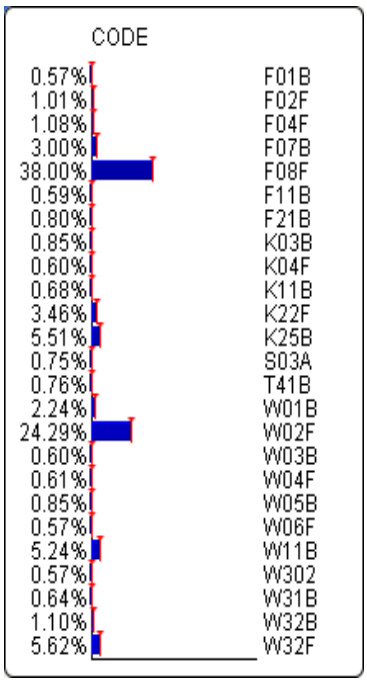
Αυτό που κάνει είναι η εκτίμηση της ποιότητας του παρόντος Μπαεζιανού δικτύου για την πρόβλεψη της ***target node*** σύμφωνα με τη συνεργαζόμενη βάση δεδομένων.Από τον "***confusion matrix***" η αξιοπιστία και η ακρίβεια των προβλέψεων κυμαίνεται σε πολύ υψηλά επίπεδα. Η μορφή του δικτύου μας είναι:

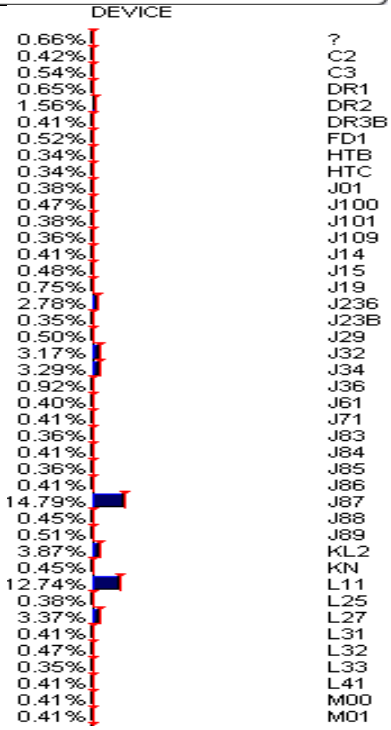


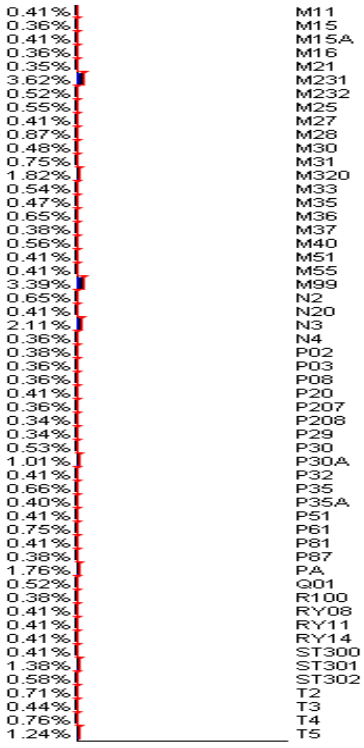
Οι προβλέψεις που δίνει είναι:











Ο διαφορετικός χρωματισμός οφείλεται στο χαρακτηρισμό της ***Source*** ως ***target***.

5.2 Εξαγωγή συμπερασμάτων

Μια επιλογή που βοηθάει στη εξαγωγή συμπερασμάτων είναι η κατάταξη των μεταβλητών ανάλογα με την «ποσότητα» της πληροφορίας που δίνει η κάθε μεταβλητή στην target node**(adaptive questionnaire based on the target).**Η μεταβλητή που δίνει τις περισσότερες πληροφορίες στην ***target*** και άρα είναι περισσότερο συσχετισμένες είναι η μεταβλητή ***Code(1.000),***έπειτα η μεταβλητή ***Device(0.8659)*** και τέλος η ***QCS(0.5381).***

Αναλυτικότερη αναφορά για τις πιθανότητες τωνν καταστάσεων της ***target node*** μπορεί να πραγματοποιηθεί μέσω της επιλογής της ***target analysis report*** .Πιο συγκεκριμένα παρατηρούμε ότι για την κατάσταση ***608*** (εμφανίζεται σε ποσοστό ***43,4%***) ,η πιο στενά συσχετιζόμενη μεταβλητή είναι η ***Code*** με την κατάσταση ***L11*** με βάρος την μονάδα.Αντίστοιχα για την ***209(28,9%)*** η σχετικότερη κατάσταση είναι η ***J87*** με βάρος την μονάδα.Η ***K25B*** (***Code***)είναι η στενότερη κατάσταση της καταστάσεως ***GT(10,6%)*** .Για την κατάσταση ***120*** η πιο στενά συνδεδεμένη μ'αυτή είναι η κατάσταση ***W11B***(***Code***) ,ενώ για την ***237(4,98%)*** αυτή είναι η κατάσταση ***101*** (***QCS***) .Για την ***910(2,28%)*** η κατάσταση που τις παρέχει τις περισσότερες πληροφορίες είναι η κατάσταση ***102***(***QCS***),και τέλος για την κατάσταση ***402***(***1,55%***) αυτό τον ρόλο τον έχει η κατάσταση ***W01B***(***Code)***

Στην προσπάθειά μας για ανίχνευση κανόνων μπορούμε να πειραματιστούμε με τις κατανομές των μεταβλητών και πόσο αλλάζουν (***propagation***).Για παράδειγμα έστω ότι η κατάσταση ***209*** της ***Source*** είναι η αιτία των προβλημάτων με πιθανότητα ***1,***τότε μπορούμε να παρατηρήσουμε ότι η πιθανότητα καταγραφής του κωδικού σφάλματος ***F08F*** είναι ***0,7954*** δηλαδή είχαμε υπερδιπλασιασμό της πιθανότητάς του, καθώς και ότι παρουσιάζεται στο εξάρτημα ***J87*** ποσοστό ***49,60%*** ενώ πριν ήταν ***14,79%.*** Εφαρμόζοντας ανάλογη διαδικασία ,θέτουμε την κατάσταση ***120*** της μεταβλητής ***Source*** εμφανίζεται με πιθανότητα ***1*** ,τότε οι πιο αξιοσημείωτες μεταβολές είναι για την κατάσταση ***237*** (της ***QCS)*** εμφανίζεται με ***0,9136*** ενώ πριν ήταν ***0,656*** και για την ***W11B***(της ***QCS***) με ***0,549*** ενώ πριν την ανανέωση των κατανομών ήταν ***0,0524.***Για την ***237*** τα αποτελέσματα που εμφανίστηκαν είναι:για την κατάσταση ***101***(της ***QCS***) με πιθανότητα ***0,9216*** (από ***0,0705***)και ο διπλασιασμός της πιθανότητας της κατάστασης ***W02F***(της ***CODE)*** από ***0,2429*** σε ***0,500***.Συνεχίζοντας με την κατάσταση ***403***(της ***Source***) ,παρατηρούμε ότι η πιθανότητα της κατάστασης ***237(***της ***QCS***) αυξήθηκε σε ***0,7222*** καθώς και για την κατάσταση ***W01B***(της ***Code***) σε ***0,2564*** από ***0,0224***. Μετατρέποντας την πιθανότητα της κατάστασης ***910***(της ***Source***) σε μονάδα,παρατηρούμε ότι η πιθανότητα της κατάστασης ***102***(της ***QCS***) αυξήθηκε σε ***0,8*** από ***0,0252***,ενώ μια σημαντκή αύξηση είχαμε στη πιθανότητα της κατάστασης ***F04F 0,2391*** από ***0,0108*** που ήταν πριν.

Τώρα θα θεωρούμε κάθε φορά γνωστή τον σταθμό καταγραφής γνωστό (δηλαδή τις καταστάσεις της μεταβλητής ***QCS***) με πιθανότητα ***1*** και θέλουμε να διαπιστώσουμε πόσο επιρρεάζει τις κατανομές των υπόλοιπων καταστάσεων των μεταβλητών και κατά συνέπεια και αν υπάρχει κάποια ενδεχομένη συσχέτιση μεταξύ τους.Έτσι μετατρέποντας την πιθανότητα της καταστάσεως ***910*** σε μονάδα ,έχουμε ότι η καινούργια πιθανότητα της καταστάσεως ***608*** (της ***Source***) είναι ***0,8262*** από ***0,4341*** και για την ***W02F 0,401*** από ***0,2429***.Για την κατάσταση ***102*** το πιο

αξιοσημείωτο αποτελέσμα ήταν για την ***910***(της ***Source***) που έγινε ***0,7247*** από ***0,0228*** ,γεγονός που φανερώνει σημαντική εξάρτηση αυτών των δύο καταστάσεων.Για την κατάσταση ***101*** η πιο σημαντική παρατήρηση που κάναμε ήταν για την πιθανότητα της καταστάσεως της ***Source*** μετατράπηκε σε ***0,6517*** από ***0,0498*** που ήταν πριν από την νέα πληρφόρηση του δικτύου .

Συνεχίζουμε θεωρώντας γνωστούς τους κωδικούς των ελαττωμάτων(δηλαδή τις καταστάσεις της ***Code***).Οι πιο εμφανείς αλλαγές παρατηρήθηκαν στις ακόλουθες περιπτώσεις:Για την κατάσταση ***F07B*** είχαμε ότι η πιθανότητα της καταστάσεως ***209***(της ***Source***) έγινε ***0,8599*** από ***0,2897***,ενώ η πιθανότητα της ***237***(της ***QCS***) μετατράπηκε σε ***0,9125*** από ***0,656*** και τέλος για τη ***J87***(της ***Device***) άλλαξε σε ***0,4274*** από ***0,1479***.Οι σημαντικότερες μεταβολές της ***K22F*** παρατηρήθηκαν στην πιθανότητα της ***GT*** που έγινε ***0,8753*** από ***0,107*** και στην ***KL2*** που από την αρχική ***0,0387*** αυξήθηκε σε ***0,2739***.Για την ***K25B*** πήραμε ότι η πιθανότητα της ***GT***(της ***Source***) έγινε ***0,8864***.Για την ***W02F*** πήραμε ότι η πιθανότητα της κατάστασης ***608(***της ***Source***) έγινε ***0,8573***.Για την ***W11B*** είχαμε ότι η πιθανότητα της κατάστασης ***120(***της ***Source***) έγινε ***0,8484*** από ***0,081***,ενώ είχαμε ότι η πιθανότητα της ***237***(της μεταβλητής ***QCS***) έγινε ***0,8530***.Τέλος για την ***W32F*** πήραμε ότι η πιθανότητα της ***608(***κατάσταση της ***Source***) έγινε ***0,9086*** από ***0,4341***.

Απ'αυτές τις παραπάνω μεταβολές μπορούμε να συμπεράνουμε ότι σε μερικές περιπτώσεις υπάρχει μεγάλη εξάρτηση μεταξύ των καταστάσεων των μεταβλητών. Εφαρμόζοντας ανάλογη διαδικασία οι υπεύθυνοι της εταιρείας θα είναι σε θέση να καταλήξουν σε πολύτιμα συμπεράσματα/κανόνες ,όπου θα βοηθήσουν στη βελτίωση της γραμμής παραγωγής ,πράγμα που δείχνει την ουσιαστικό ρόλο που διαδραμάτισε η εφαρμογή των ***Bayesian Networks***

Παράρτημα Α

Πίνακας Ι: Επεξήγηση κωδικών ελαττωμάτων

Σταθμός εργασίας	Ελαττώματα		
	Κωδικός	Επεξήγηση	Πηγή Χρέωσης
Q1	K04F	Ελαττωματικό SMD Εξάρτημα	P1
Οπτικός SMD	K22F	Σπασμένο SMD εξάρτημα	FM
	W02F	Απουσία SMD Εξαρτήματος	
	W04F	Λάθος SMD Εξάρτημα	
	W06F	Ανάστροφη Τοποθέτηση SMD υλικών	
	W30K	Λάθος Θέση SMD εξαρτήματος	
	W32B	Κακή τοποθέτηση SMD εξαρτήματος	
	F02F	Απουσία κόλλησης στη πλευρά SMD εξαρτημάτων	
	F04F	Βραχυκυκλώματα από κόλληση σε υλικά SMD	
	F08F	Ακόλλητο SMD εξάρτημα	
	F22F	Υπολείμματα Κόλλησης	

Ελαττώματα			
Σταθμός εργασίας	**Κωδικός**	**Επεξήγηση**	**Πηγή Χρέωσης**
Q2	K01B	Ελαττωματική πλακέτα	P1
Οπτικός	K03B	Ελαττωματικό Εξάρτημα	P2
	K04F	Ελαττωματικό SMD Εξάρτημα	P3
	K09B	Εξάρτημα εκτός ανοχών	P4
	K11B	Καμένο Εξάρτημα	Q1
	K21B	Σπασμένο/γδαρμένο εξάρτημα	Q2
	K22F	Σπασμένο SMD εξάρτημα	FM
	K25B	Κακή συγκολησιμότητα	
	W01B	Απουσία Εξαρτήματος	
	W02F	Απουσία SMD Εξαρτήματος	
	W03B	Λάθος Εξάρτημα	
	W04F	Λάθος SMD Εξάρτημα	
	W05B	Ανάστροφη Τοποθέτηση	
	W06F	Ανάστροφη Τοποθέτηση SMD υλικών	
	W07B	Λάθος/Απουσία γεφύρωσης	
	W11B	Μη ορατοί ακροδέκτες	
	W13B	Προεξοχή ακροδεκτών	
	W15B	Κλίση εξαρτήματος	
	W17B	Απόσταση εξαρτήματος από πλακέτα	
	W19B	Στραβωμένος/κομένος ακροδέκτης	
	W25B	Πρόκληση βραχυκυκλώματος	
	W30K	Λάθος Θέση SMD εξαρτήματος	
	W31B	Κακή τοποθέτηση εξαρτήματος	
	W32B	Κακή τοποθέτηση εξαρτήματος	
	W32F	Κακή τοποθέτηση SMD εξαρτήματος	
	T31B	Δυσλειτουργία μηχανής/συσκευής	
	F01B	Απουσία κόλλησης στη πλευρά εξαρτημάτων	
	F02F	Απουσία κόλλησης στη πλευρά SMD εξαρτημάτων	
	F03B	Βραχυκυκλώματα από κόλληση	
	F04F	Βραχυκυκλώματα από κόλληση σε υλικά SMD	
	F05B	Υπερβολική/λίγη ποσότητα κόλλησης	
	F07B	Ακόλλητο εξάρτημα	
	F08F	Ακόλλητο SMD εξάρτημα	
	F10B	Κρατήρες πάνω από 1% των κολλήσεων	
	F11B	Ψυχρή κόλληση	
	F13B	Ακόλλητη επιφάνεια	
	F15B	Υπολείμματα flux	
	F17B	Ξένες ουσίες	

Ελαττώματα

Σταθμός εργασίας	**Κωδικός**	**Επεξήγηση**	**Πηγή Χρέωσης**
Q3	K03B	Ελαττωματικό Εξάρτημα	P1
Λειτουργικός έλεγχος	K04F	Ελαττωματικό SMD Εξάρτημα	P2
	K09B	Εξάρτημα εκτός ανοχών	P3
	K11B	Καμένο Εξάρτημα	P4
	K13B	Κακή κατάσταση μόνωσης καλωδίου	Q1
	K21B	Σπασμένο/γδαρμένο εξάρτημα	Q2
	K22F	Σπασμένο SMD εξάρτημα	FM
	K25B	Κακή συγκολησιμότητα	
	W01B	Απουσία Εξαρτήματος	
	W03B	Λάθος Εξάρτημα	
	W05B	Ανάστροφη Τοποθέτηση	
	W11B	Μη ορατοί ακροδέκτες	
	W19B	Στραβομένος/κομένος ακροδέκτης	
	W25B	Πρόκληση βραχυκυκλώματος	
	W31B	Κακή τοποθέτηση εξαρτήματος	
	W32F	Κακή τοποθέτηση SMD εξαρτήματος	
	D01B	Ελαττωματικό PBA	
	F11B	Ψυχρή κόλληση	

Ελαττώματα			
θμός εργασίας	**Κωδικός**	**Επεξήγηση**	**Πηγή Χρέωσης**
Q4	K03B	Ελαττωματικό Εξάρτημα	P1
Burn In – Τελικός	K04F	Ελαττωματικό SMD Εξάρτημα	P2
	K09B	Εξάρτημα εκτός ανοχών	P4
	K11B	Καμένο Εξάρτημα	P3
	K13B	Κακή κατάσταση μόνωσης καλωδίου	Q1
	K21B	Σπασμένο/γδαρμένο εξάρτημα	Q2
	K22F	Σπασμένο SMD εξάρτημα	FM
	W02K	Απουσία SMD Εξαρτήματος	
	W04K	Λάθος SMD Εξάρτημα	
	W06F	Ανάστροφη Τοποθέτηση SMD υλικών	
	W19B	Στραβομένος/κομένος ακροδέκτης	
	W25B	Πρόκληση βραχυκυκλώματος	
	W30F	Λάθος Θέση SMD εξαρτήματος	
	W32F	Κακή τοποθέτηση SMD εξαρτήματος	
	D01B	Ελαττωματικό PBA	
	T31B	Δυσλειτουργία μηχανής/συσκευής	
	F01B	Απουσία κόλλησης στη πλευρά εξαρτημάτων	
	F02F	Απουσία κόλλησης στη πλευρά SMD εξαρτημάτων	
	F03B	Βραχυκυκλώματα από κόλληση	
	F04F	Βραχυκυκλώματα από κόλληση σε υλικά SMD	
	F07B	Ακόλλητο εξάρτημα	
	F08F	Ακόλλητο SMD εξάρτημα	
	F11B	Ψυχρή κόλληση	
	F21B	Υπολείμματα κόλλησης	
	K03B	Ελαττωματικό Εξάρτημα	

Παράρτημα Β

Πίνακας επεξήγησης των συμβολισμών

Αυτή η λίστα αποτυπώνει τα πιο σημαντικά σύμβολα και συντήσεις που χρησιμοποιήθηκαν σε αυτό το κεφάλαιο με στόχο την καλύτερη κατανόηση της θεωρίας των *Bayesian Networks*

Μεταβλητές

α,b,x_i,υ,...	**μεταβλητές**
S,…	**σύνολο μεταβλητών**
|S|	**αριθμός των μεταβλητών,που επίσης καλείται και "*order* "ενός συνόλου *S***
$\mathcal{V}$	**σύνολο όλων των μεταβλητών στο domain**
I(α)	**σύνολο όλων των καταστλασεων της μεταβλητής α**
I(S)	**σύνολο των κοινών καταστάσεων των μεταβλητών του S**
i,j,… $\in$ ***I(S)***	**μια κοινή κατάσταση των μεταβλητών στο *S***

Μπαεζιανά δίκτυα,γραφήματα και παράμετροι

θ	**παράμετροι ενός Μπαεζιανού δικτύου**
m	**directed acyclic graph (DAG)**
$\bar{m}$	**skeleton**
M	**σύνολο όλων των DAGs**
α~b	**(απροσανατολισμένο) τόξο μεταξύ των μεταβλητών α και b**
α→b, b←α	**προσανατολισμένο τόξο μεταξύ των μεταβλητών α και b**
pa_m(α),pa(α)	**γονείς μιας μεταβλητής $a \in \mathcal{V}$ στο *DAGm***
an(α)	**πρόγονοι μιας μεταβλητής $a \in \mathcal{V}$ στο *DAG***
ne(α)	**γείτονες μιας μεταβλητής $a \in \mathcal{V}$ στο γράφημα**
$a \not\perp b \mid \mathcal{S}$	**d-connection των *α* και*b* δεδομένου του *S***
$a \perp b \mid \mathcal{S}$	**d-separation των *α* και*b* δεδομένου του *S***

Πιθανότητες και Ανεξαρτησίες

p(α) **(οριακή) κατανομή πιθανότητας για τη μεταβλητή *α***
p(α,b) ,p(S) **κοινή κατανομή πιθανότητας για διάφορες μεταβλητές**
p(α|S) **κατά συνθήκη κατανομή της μεταβλητής α δεδομένου του *S***
$a \perp\!\!\!\perp b$ **οριακή ανεξαρτησία των μεταβλητών *α* και *b***
$a \not\perp\!\!\!\perp b$ **οριακή εξάρτηση των μεταβλητών *α* και *b***
$a \perp\!\!\!\perp b \mid S$ **κατά συνθήκη ανεξαρτησία των μεταβλητών *α* και *b* δεδομένου του *S***
$a \not\perp\!\!\!\perp b \mid S$ **κατά συνθήκη εξάρτηση των μεταβλητών *α* και *b* δεδομένου του *S***
CIDs **Conditional Independences and Dependences**

Βιβλιογραφία

[1]HUGIN Expert A/S. Aalborg, Denmark, http://www.hugin.com.

[2] M. I. Jordan, editor. *Learning in Graphical Models*. Kluwer Academic Publishers, The Netherlands, 1998.

[3] F. V. Jensen. *An Introducution to Bayesian Networks*. University College London Press, 1996.

[4]F.V. Jensen. *Bayesian Networks and Decision Digrams.* Springer.2001

[5] K.P. Murphy. *An introduction to graphical models*

[6] Steck, H. . *Constrained-Based Structural Learning in Bayesian Networks Using Finite Data Sets*, PhD Thesis, Institut fór der Informatik der Technischen Universitat Munchen. 2001

[7] BayesiaLab . http://www.bayesia.com

[8] S.Lauritzen. *Graphical Models* ,Oxford.1996

[9] D.Margaritis . *Learning Bayesian Network Model Structure from Data* PhD Thesis,2003

[10] R.Neapolitan. *Learning Bayesian Networks*

[11] http://www.ai.mit.edu/~murphyk/Bayes/bayes.html

[12] D. Heckerman, D. Geiger, and D. M. Chickering. *Learning Bayesian networks: The combination of knowledge and statistical data.* Machine Learning, 20:197-243, 1995a.

[13] D. Heckerman. *A tutorial on learning with Bayesian networks*. Technical Report MSR-TR-95-06, Microsoft Research, Advanced Technology Division, March 1995.

[14] D. M. Chickering. *Learning Bayesian networks is NP-complete*. In Proceedings of AI and Statistics, 1995.

[15] F. V. Jensen *Bayesian Networks and Decision Graphs,* Springer 2001

[16] SerafÌn Moral **.** *Structural Learning: The PC Algorithm*

[17] David Grannen ,Mathieu Robin ,Micheal Lynch ,Sohail Akram,Tolu Aina. *Bayesian Network*

[18] Dennis M. Buede ,Joseph A. Tatman,Terry A. Bresnick. *Introduction to Bayesian Networks* A Tutorial for the 66th MORS Symposium

[19] Steck, H. and Tresp, V. (1999). *Bayesian Belief Networks for Data Mining, Proceedings of The 2nd Workshop on Data Mining and Data Warehousing* Sammelband, Universitδt Magdeburg, September 1999

[20] Λουκάς Τσιρώνης .*Εμπλουτισμός της διαδικασίας ολικής ποιότητας με τεχνικές πρόσληψης γνώσεων* ,Διδακτορική εργασία