ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Bayesian δίκτυα και εύρεση δομής με ακριβή αλγόριθμο

Μάρκος Ν. Αρβανίτης

Επιβλέποντες: **Ιωάννης Ιωαννίδης,** Καθηγητής
Όμηρος Μεταξάς, Ερευνητής

ΑΘΗΝΑ

ΣΕΠΤΕΜΒΡΙΟΣ 2015

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Bayesian δίκτυα και εύρεση δομής με ακριβή αλγόριθμο

Μάρκος Ν. Αρβανίτης
Α.Μ.: 1115200900045

ΕΠΙΒΛΕΠΟΝΤΕΣ: **Ιωάννης Ιωαννίδης,** Καθηγητής
Όμηρος Μεταξάς, Ερευνητής

ΠΕΡΙΛΗΨΗ

Η εύρεση δομής σε δίκτυα Bayesian έχει κεντρίσει το ενδιαφέρον πολλών ερευνών την τελευταία δεκαετία. Οι εφαρμογές τους ποικίλλουν, από τον αθλητισμό, την ψυχαγωγία μέχρι και την ιατρική. Ο σκοπός της εργασίας αυτής είναι να παρουσιάσει τα βασικά χαρακτηριστικά των Bayesian δικτύων, να αναλύσει τη δομή τους και να επικεντρωθεί στο πρόβλημα της εύρεσης αυτής της δομής. Συγκρίνονται διαφορετικές μέθοδοι για την επίλυση του προβλήματος αυτού και υλοποιείται ένας ακριβής αλγόριθμος για την εύρεση της δομής τέτοιων δικτύων μέσα από δεδομένα που παρέχονται. Στόχος της εργασίας είναι επίσης η ενσωμάτωση αυτού του αλγορίθμου στο πρόγραμμα AITION. Τέλος, εκτελούνται πειραματικές μελέτες με σκοπό την εξακρίβωση της αποδοτικότητας του αλγορίθμου αυτού και παρουσιάζονται τα αποτελέσματα τους.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Bayesian δίκτυα

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: εύρεση δομής, συμπερασμός, δυναμικός προγραμματισμός, ακριβής αλγόριθμος, AITION

ABSTRACT

Structure discovery has been the main focus on multiple research papers during the last decade. It can be applied in many fields including sports, entertainment and even medical science. The purpose of this paper is to present the basic characteristics of Bayesian networks, to analyse their structure and to emphasise on the problem of discovering this structure. Various different solutions are compared and an exact algorithm for Bayesian structure discovery through data is implemented. In addition to this, a focal point is the incorporation of this algorithm in the AITION platform. Finally, experiments are conducted to prove the efficiency of the algorithm and their results presented thoroughly.

SUBJECT AREA: Bayesian Networks

KEYWORDS: structure discovery, inference, dynamic programming, exact algorithm, AITION

ΕΥΧΑΡΙΣΤΙΕΣ

Αρχικά θα ήθελα να εκφράσω ένα μεγάλο ευχαριστώ στον επιβλέποντα καθηγητή μου κ. Ιωάννη Ιωαννίδη για την καθοδήγηση και το ενδιαφέρον του καθώς επίσης και για την υποστήριξή του. Επίσης θα ήθελα να ευχαριστήσω τον κ. Όμηρο Μεταξά, που βρισκόταν συνεχώς παρών στην πορεία της πτυχιακής, πάντα πρόθυμος για να με καθοδηγήσει σε όποια δυσκολία συναντούσα και να με παροτρύνει για την ολοκλήρωση της εργασίας. Το ενδιαφέρον τους για το αντικείμενο συνέβαλλε πολύ στην εκπόνηση ενός καλού αποτελέσματος.

Περιεχόμενα

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΠΡΟΛΟΓΟΣ

Η παρούσα πτυχιακή εργασία με τίτλο «Bayesian δίκτυα και εύρεση δομής με ακριβή αλγόριθμο» εκπονήθηκε στο πλαίσιο του Προπτυχιακού προγράμματος, του τμήματος Πληροφορικής και Τηλεπικοινωνιών του Εθνικού Καποδιστριακού Πανεπιστημίου Αθηνών.

1. BAYESIAN ΔΙΚΤΥΑ

1.1 Εισαγωγή στα Bayesian Δίκτυα

Ένα Bayesian δίκτυο[1] είναι ένα γραφικό μοντέλο που κωδικοποιεί πιθανοτικές σχέσεις ανάμεσα σε ένα σύνολο μεταβλητών. Τις προηγούμενες δεκαετίες το Bayesian δίκτυο έχει εξελιχθεί σε μια δημοφιλή αναπαράσταση για την κωδικοποίηση αβέβαιης ειδικής γνώσης σε έμπειρα συστήματα. Τα τελευταία χρόνια ερευνητές ανέπτυξαν μεθόδους για την εκμάθηση Bayesian δικτύων από δεδομένα. Οι τεχνικές που αναπτύχθηκαν είναι σχετικά καινούριες και εξελίσσονται ακόμα, αλλά έχει αποδειχθεί ότι είναι αξιοσημείωτα αποδοτικές σε μερικά προβλήματα ανάλυσης δεδομένων.

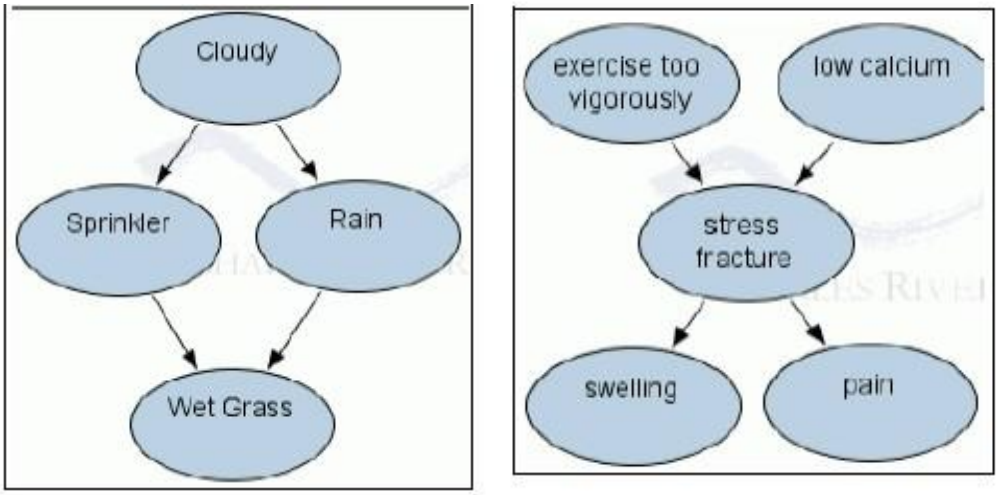
Επομένως, τι προσφέρουν τα Bayesian δίκτυα και οι Bayesian μέθοδοι; Αρχικά τα Bayesian δίκτυα μπορούν να διαχειριστούν εύκολα ελλιπή σύνολα δεδομένων. Για παράδειγμα ας θεωρήσουμε ένα πρόβλημα ταξινόμησης, όπου δύο από τις μεταβλητές εισόδου είναι αυστηρώς αντί-συσχετιζόμενες. Ο συσχετισμός αυτών των μεταβλητών δεν αποτελεί δύσκολο πρόβλημα αφού υπάρχουν πρότυπες επιβλεπόμενες τεχνικές μάθησης κατά τις οποίες εκτιμώνται όλες οι είσοδοι. Ωστόσο σε περιπτώσεις όπου δεν μπορεί να εκφραστεί μία από τις μεταβλητές εισόδου τα περισσότερα μοντέλα παράγουν ελλιπείς προβλέψεις αφού δεν μπορούν να κωδικοποιήσουν τις συσχετίσεις ανάμεσα στις μεταβλητές αυτές. Τα Bayesian δίκτυα παρέχουν φυσικές μεθόδους για την κωδικοποίηση τέτοιων εξαρτήσεων. Κατά δεύτερον τα Bayesian δίκτυα επιτρέπουν στην εκμάθηση των τυχαίων σχέσεων των μεταβλητών, γεγονός που θεωρείται σημαντικό για δύο λόγους. Αρχικά η διαδικασία είναι χρήσιμη κατά την προσπάθεια κατανόησης μιας προβληματικής περιοχής, για παράδειγμα κατά την επεξηγηματική ανάλυση δεδομένων. Επιπρόσθετα η γνώση τυχαίων σχέσεων επιτρέπει την πρόβλεψη παρεμβάσεων. Για παράδειγμα, ένας αναλυτής εμπορίου χρειάζεται να ξέρει εάν αξίζει ή όχι η έκθεση μιας συγκεκριμένης διαφήμισης με σκοπό την αύξηση των πωλήσεων ενός προϊόντος. Για την απάντηση της ερώτησης, ο αναλυτής μπορεί να προσδιορίσει εάν η διαφήμιση οδηγεί στην αύξηση των πωλήσεων και σε ποιό βαθμό. Η χρήση των Bayesian δικτύων βοηθά στο να απαντηθούν τέτοια ερωτήματα, ακόμα και αν δεν υπάρχουν έρευνες και πειράματα σχετικά με τις συνέπειες της αύξησης της έκθεσης των προϊόντων μέσω διαφήμισης.

Τα Bayesian δίκτυα μαζί με τις Bayesian στατιστικές τεχνικές προωθούν τον συνδυασμό της πρότερης γνώσης και των δεδομένων. Είναι γενικά αποδεκτή η αξία της πρότερης γνώσης ειδικά όταν τα δεδομένα είναι σπάνια και ακριβά. Το γεγονός ότι μερικά εμπορικά συστήματα (έμπειρα συστήματα) μπορούν να δομηθούν βασισμένα αποκλειστικά σε πρότερη γνώση αποτελεί διαθήκη της αξίας της γνώσης αυτής. Τα Bayesian δίκτυα υποστηρίζουν την τυχαία σημασιολογία, με αποτέλεσμα να επιτρέπουν την άμεση κωδικοποίηση της τυχαίας πρότερης γνώσης. Επιπλέον κωδικοποιούν τις τυχαίες σχέσεις με πιθανότητες. Επομένως, η πρότερη γνώση και τα δεδομένα μπορούν να συνδυαστούν με τεχνικές της Bayesian στατιστικής.

1.2 Βασικά Χαρακτηριστικά των δικτύων Bayes

Τα δίκτυα Bayes αποτελούν ισχυρά εργαλεία για τη μοντελοποίηση αιτίας-συνέπειας σε μια ευρεία ποικιλία εφαρμογών. Αποτελούν συμπαγή δίκτυα πιθανοτήτων, που υιοθετούν πιθανοτικές σχέσεις μεταξύ των μεταβλητών καθώς επίσης και ιστορικές πληροφορίες των σχέσεων αυτών.

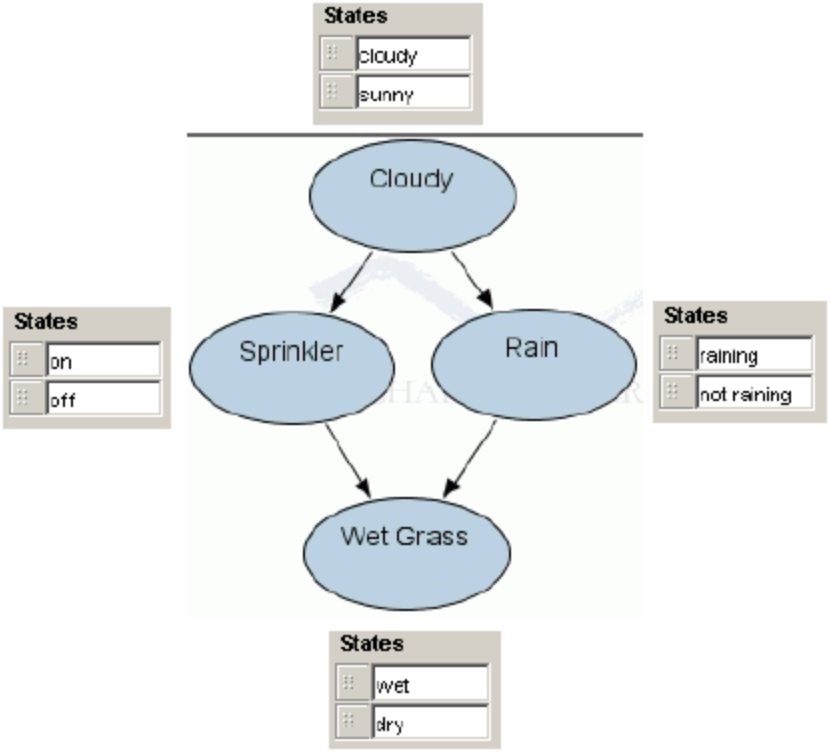
Τα Bayesian δίκτυα, όπως προαναφέρθηκε, είναι πολύ αποδοτικά για τη μοντελοποίηση καταστάσεων όπου κάποια πληροφορία είναι ήδη γνωστή και τα εισερχόμενα δεδομένα δεν είναι συγκεκριμένα ή δεν είναι διαθέσιμα. Αυτά τα δίκτυα παρέχουν επίσης λογικές σημασιολογίες για την αναπαράσταση των αιτιών και συνεπειών, καθώς και πιθανοτήτων μέσω μιας διαισθητικής γραφικής αναπαράστασης.



Εικόνα 1 Παραδείγματα Bayesian δικτύων

Με απλούς όρους, τα Bayesian δίκτυα, είναι μοντέλα που μπορούν να μοντελοποιήσουν οποιαδήποτε κατάσταση: τον καιρό, ένα στρατιωτικό τάγμα, μία ασθένεια και τα συμπτώματά της, κοκ. Επιπλέον είναι ιδιαίτερα χρήσιμα όταν η πληροφορία σχετικά με το παρελθόν και την τρέχουσα κατάσταση είναι ατελής, ασαφής και αόριστη.

Κάθε μεταβλητή σε ένα Bayesian δίκτυο αναπαριστάται από κόμβους (nodes) (Εικόνα 1). Μια μεταβλητή μπορεί να αναπαριστά έναν διακόπτη φωτός ο οποίος μπορεί να είναι είτε ανοιχτός είτε κλειστός, την εγγύτητα ενός εχθρικού τάγματος, τον καιρό μιας περιοχής, κα. Κάθε κόμβος χαρακτηρίζεται από καταστάσεις (states), δηλαδή χαρακτηρίζεται από ένα σύνολο από πιθανές τιμές που αντιστοιχούν σε κάθε μεταβλητή. Για παράδειγμα ο καιρός μπορεί να είναι συννεφιασμένος, βροχερός ή ηλιόλουστος, ένα εχθρικό τάγμα μπορεί να είναι κοντά ή μακριά, τα συμπτώματα μιας ασθένειας μπορούν είτε να είναι εμφανή είτε όχι κλπ. Οι κόμβοι συνδέονται μεταξύ τους με κατευθυνόμενα βέλη (ή πλευρές) (edges). Τα βέλη αυτά φανερώνουν την αλληλεξάρτηση των μεταβλητών υποδεικνύοντας και την κατεύθυνση της επιρροής.



Εικόνα 2 Κόμβοι, πλευρές και καταστάσεις σε ένα Bayesian δίκτυο.

Στο απλό μοντέλο της εικόνας 2 ο καιρός μπορεί να είναι είτε συννεφιασμένος είτε ηλιόλουστος. Το γεγονός ότι βρέχει ή όχι εξαρτάται από τη συννεφιά. Επιπλέον το γρασίδι μπορεί να είναι βρεγμένο ή στεγνό και ο μηχανισμός αυτόματου ποτίσματος μπορεί να είναι ανοιχτός ή κλειστός. Υπάρχει επιπλέον ακόμα μία αλληλεξάρτηση. Εάν ο καιρός είναι βροχερός τότε το γρασίδι θα βραχεί αμέσως. Ωστόσο και σε ηλιόλουστο καιρό μπορεί να παρατηρηθεί βρεγμένο γρασίδι σε περίπτωση που ο ιδιοκτήτης ανοίξει το μηχάνημα του αυτόματου ποτίσματος.

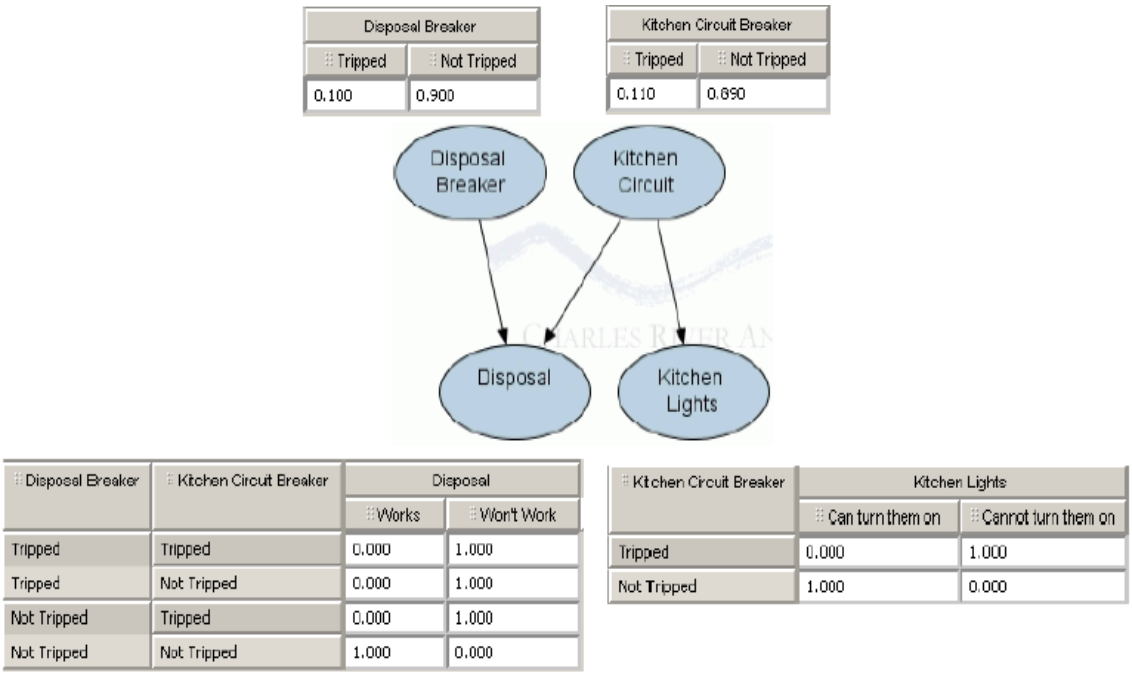
Ένα Bayesian δίκτυο είναι ένα μοντέλο που αναπαριστά τις πιθανές καταστάσεις μιας δοθείσας θεματικής περιοχής. Περιέχει επιπλέον πιθανοτικές σχέσεις ανάμεσα σε ορισμένες καταστάσεις της περιοχής αυτής. Για παράδειγμα όταν εισάγονται πιθανότητες στο Bayesian δίκτυο το οποίο αναπαριστά τον καιρό και τη χρήση του μηχανισμού αυτόματου ποτίσματος, τότε αυτό το δίκτυο μπορεί να χρησιμοποιηθεί για να απαντηθούν ερωτήσεις όπως οι παρακάτω:

- Εάν το γρασίδι είναι βρεγμένο αυτό προήλθε από τη βροχή ή από τον μηχανισμό αυτόματου ποτίσματος;
- Πόσο πιθανό είναι να πρέπει να ποτίσουμε το γρασίδι μια συννεφιασμένη μέρα;

Η πιθανότητα οποιουδήποτε κόμβου του δικτύου να βρίσκεται σε μια κατάσταση χωρίς να υπάρχει κάποια τρέχουσα ένδειξη περιγράφεται

χρησιμοποιώντας έναν πίνακα υπό συνθήκη πιθανοτήτων. Οι πιθανότητες σε κάποιους κόμβους επηρεάζονται από την κατάσταση κάποιων άλλων κόμβων, ανάλογα με τις αλληλεξαρτήσεις που υπάρχουν. Η πρότερη πληροφορία σχετικά με τις σχέσεις ανάμεσα στους κόμβους υποδεικνύει σε ορισμένες περιπτώσεις την πιθανότητα του να βρίσκεται ένας κόμβος σε μία κατάσταση ανάλογα με την κατάσταση ενός άλλου κόμβου.

Για παράδειγμα, πρότερη πληροφορία μπορεί να φανερώνει ότι εάν ο καιρός είναι συννεφιασμένος τότε η πιθανότητα να βρέξει είναι υψηλότερη. Ένα παράδειγμα των υπό συνθήκη πιθανοτήτων ενός Bayesian δικτύου φαίνεται στην εικόνα 3.



Εικόνα 3 Πίνακες υπό συνθήκη πιθανοτήτων για κάποιο Bayesian δίκτυο

Με την πρότερη πληροφορία αποθηκευμένη στον πίνακα των υπό συνθήκη πιθανοτήτων, τα δίκτυα Bayes μπορούν να βοηθήσουν στο να πάρουμε κάποια απόφαση, ή ως ένας τρόπος αυτοματοποίησης μιας διαδικασίας διαμόρφωσης απόφασης. Επιπλέον μπορούν να χρησιμοποιηθούν στην εκτέλεση επαγωγικού συλλογισμού, δηλαδή την διάγνωση αιτίας δοθέντων κάποιων συνεπειών αλλά και το ανάποδο, παραγωγικού συλλογισμού, δηλαδή την πρόβλεψη συνεπειών δοθείσας της αιτίας.

1.3 Κανόνας του Bayes

Τα Bayesian δίκτυα βασίζονται στην εργασία του μαθηματικού **Thomas Bayes**, ο οποίος ασχολήθηκε με την υπό συνθήκη πιθανοτική θεωρία στα

τέλη του 1700 και ανακάλυψε ένα βασικό νόμο πιθανοτήτων ο οποίος αργότερα ονομάστηκε κανόνας του Bayes.

Με απλό τρόπο ο κανόνας του Bayes μπορεί να εκφραστεί ως εξής:

$$P(b|a) = \frac{P(a|b) \times P(b)}{P(a)}$$

Όπου P(a) είναι η πιθανότητα του a και P(a|b) είναι η πιθανότητα του a δοθέντος ότι το b έχει συμβεί.

Για παράδειγμα ας υποθέσουμε ότι η μηνιγγίτιδα μπορεί να προκαλέσει δύσκαμπτο λαιμό σε ποσοστό 50%. Επιπλέον υποθέτουμε ότι γνωρίζουμε από πληθυσμιακές μελέτες ότι 1/ 50000 ανθρώπους έχουν μηνιγγίτιδα και 1/ 20 έχει δύσκαμπτο λαιμό. Θέλουμε να βρούμε την πιθανότητα ένας ασθενής που παραπονιέται για δύσκαμπτο λαιμό, να έχει μηνιγγίτιδα. Αναλυτικότερα πόσο πιθανή είναι η μηνιγγίτιδα δεδομένου ότι υπάρχει δύσκαμπτος λαιμός.

Για να απεικονίσουμε το παραπάνω υπολογίζουμε:

P(μηνιγγίτιδα δύσκαμπτος λαιμός) =

P(δύσκαμπτος λαιμός| μηνιγγίτιδα) *P(μηνιγγίτιδα) /P(δύσκαμπτος λαιμός)=

$$= \frac{0.5 \times 1/50{,}000}{1/20} = 0.0002$$

Επομένως εάν ένας ασθενής παραπονιέται ότι έχει δύσκαμπτο λαιμό τότε η πιθανότητα αυτό να οφείλεται από μηνιγγίτιδα είναι μόνο 0.0002.

Ένας περισσότερο πολύπλοκος τρόπος έκφρασης του κανόνα του Bayes, ο οποίος περιλαμβάνει υπόθεση, πρότερη εμπειρία και ένδειξη είναι ο εξής:

$$P(H|E, c) = \frac{P(H|c) \times P(E|H,c)}{P(E|c)}$$

Με την παραπάνω σχέση μπορούμε να ανανεώσουμε την πεποίθησή μας για την υπόθεση H δοθείσας της πρόσθετης ένδειξης E και της πρότερης εμπειρίας c.

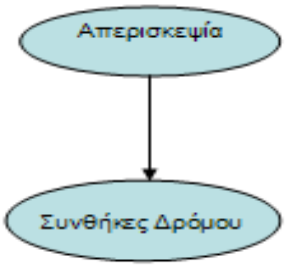
Ο αριστερός όρος P (H|E,c) ονομάζεται μεταγενέστερη (posterior) πιθανότητα ή αλλιώς πιθανότητα της υπόθεσης H αφού λάβουμε υπόψη τη συνέπεια της ένδειξης E στην πρότερη εμπειρία c. Ο όρος P (H|c) καλείται εκ των προτέρων (a- priori) πιθανότητα της H δοθείσας μόνο της c. Ο όρος P (E|H,c) καλείται πιθανότητα (likelihood) και δίνει την πιθανότητα της ένδειξης αν δεχτούμε ότι η υπόθεση H και η πρότερη πληροφορία c είναι αληθείς (true). Τέλος ο όρος P (E|c) είναι ανεξάρτητος του H και μπορεί να θεωρηθεί ως παράγοντας κανονικοποίησης ή κλιμάκωσης. Τα Bayesian δίκτυα υιοθετούν τον κανόνα του Bayes σε ένα γραφικό μοντέλο.

1.4 Δομή των Bayesian δικτύων

Στη συγκεκριμένη παράγραφο θα δούμε αναλυτικότερα τα βασικά δομικά χαρακτηριστικά των Bayesian δικτύων [2]. Όπως είδαμε παραπάνω τα Bayesian δίκτυα, γραφικά, είναι μοντέλα όπου κάθε μεταβλητή αναπαριστάται με ένα κόμβο, και οι σχέσεις αλληλεξάρτησης σημειώνονται με βέλη- πλευρές.

1.4.1 Κόμβοι (Nodes)

Ένας κόμβος αποτελεί την αναπαράσταση μιας μεταβλητής στην κατάσταση την οποία έχει μοντελοποιηθεί. Ένας κόμβος αποτυπώνεται γραφικά με ένα οβάλ, το οποίο χαρακτηρίζεται από κάποια ετικέτα. Το απλό παράδειγμα στο σχήμα 5.4 δείχνει δύο κόμβους, απ τους οποίους ο ένας φέρει την ετικέτα «Απερισκεψία» και ο άλλος χαρακτηρίζεται ως «Συνθήκες Δρόμου».



Εικόνα 4 Δύο κόμβοι και μία πλευρά σε ένα απλό Bayesian δίκτυο

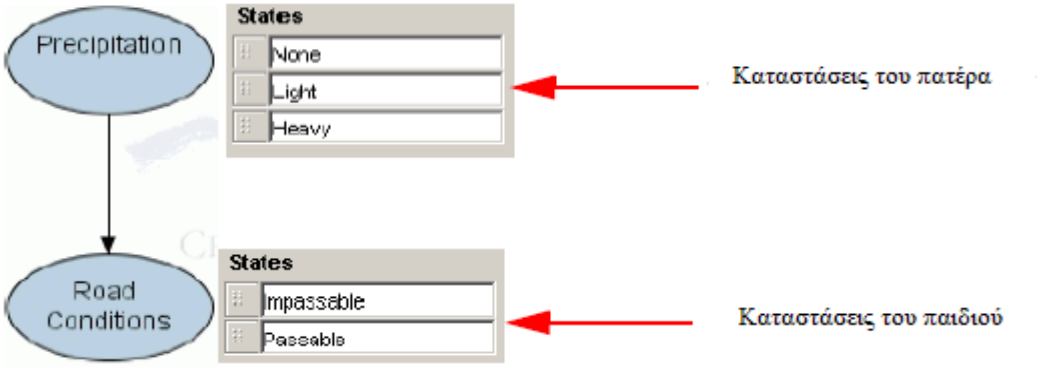
1.4.2 Πλευρές (Edges)

Μια πλευρά εμφανίζει μια σχέση αλληλεξάρτησης ανάμεσα σε δύο κόμβους. Γραφικά, απεικονίζεται με ένα κατευθυνόμενο βέλος μεταξύ των κόμβων και η κατεύθυνση αυτού υποδεικνύει την κατεύθυνση της εξάρτησης. Διαισθητικά, η σημασία μιας πλευράς που σχεδιάζεται από τον κόμβο Χ στον κόμβο Υ είναι ότι ο κόμβος Χ έχει μία άμεση επιρροή στον κόμβο Υ. Για παράδειγμα, στην εικόνα 4, η πλευρά φανερώνει ότι το επίπεδο της ‘απερισκεψίας’ επηρεάζει άμεσα τις ‘συνθήκες του δρόμου’. Ο τρόπος που ένας κόμβος επηρεάζει κάποιον άλλο καθορίζεται από τον πίνακα των υπό συνθήκη πιθανοτήτων.

Οι πλευρές καθορίζουν επίσης κάποιους χαρακτηριστικούς όρους για τους κόμβους. Όταν δύο κόμβοι συνδέονται με μια πλευρά, ο κόμβος από τον οποίο ξεκινάει το βέλος ονομάζεται πατέρας (parent) αυτού στον οποίο καταλήγει. Στο παράδειγμά μας ο κόμβος «Απερισκεψία» αποτελεί πατέρα του «Συνθήκες Δρόμου». Κατά συνέπεια ο κόμβος παιδί εξαρτάται από τον πατέρα του.

1.4.3 Καταστάσεις (States)

Οι τιμές που μπορούν να πάρουν οι μεταβλητές ονομάζονται καταστάσεις. Για παράδειγμα οι σημαντικές καταστάσεις της μεταβλητής 'Απερισκεψία' είναι οι: Καθόλου (None), Μικρή (Light) και Μεγάλη (Heavy). Επιπλέον είναι γνωστό ότι η απερισκεψία μπορεί να οδηγήσει στην διάσχιση (passable) ή όχι του δρόμου. Μπορούμε να παρατηρήσουμε τις καταστάσεις κάθε κόμβου στην εικόνα 5.



Εικόνα 5 Καταστάσεις και τιμές που μπορεί να πάρει ένας κόμβος

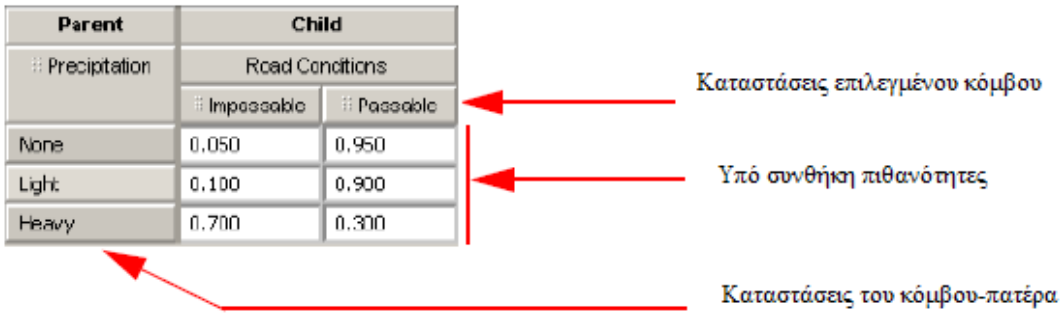
1.4.4 Πίνακες των υπό συνθήκη πιθανοτήτων

Σε κάθε κόμβο αντιστοιχεί και ένας πίνακας υπό συνθήκη πιθανοτήτων (conditional probability table). Οι υπό συνθήκη πιθανότητες εκφράζουν πιθανότητες οι οποίες βασίζονται σε πρότερη πληροφορία και παρελθοντική εμπειρία.

Η υπό συνθήκη πιθανότητα καθορίζεται μαθηματικά με τη σχέση P (x|p1,p2,..pn) και εκφράζει την πιθανότητα του να βρίσκεται η μεταβλητή X σε μια κατάσταση x εάν ο πατέρας P1 βρίσκεται στην κατάσταση p1, ο πατέρας P2 στην κατάσταση p2,..., και ο πατέρας Pn στην κατάσταση pn.

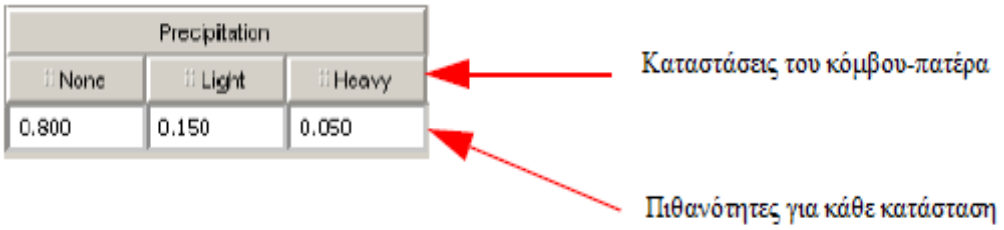
Βασισμένοι στον παραπάνω ορισμό συμπεραίνουμε ότι για κάθε πατέρα και κάθε πιθανή κατάσταση του πατέρα αυτού, υπάρχει μία γραμμή στον πίνακα πιθανοτήτων η οποία περιγράφει την πιθανότητα του να είναι ο κόμβος παιδί σε κάποια συγκεκριμένη κατάσταση. Για παράδειγμα, το πρώτο κελί του πίνακα πιθανοτήτων στην εικόνα 6, το οποίο αναφέρεται στον κόμβο «Συνθήκες Δρόμου» μπορεί να ερμηνευτεί ως εξής: Εάν ο κόμβος πατέρας «Απερισκεψία» είναι στην κατάσταση «Καθόλου» τότε η πιθανότητα του να είναι ο κόμβος «Συνθήκες Δρόμου» στην κατάσταση αδιάβατος είναι 5%. Αντιστοίχως μπορεί να ερμηνευτεί κάθε κελί του πίνακα πιθανοτήτων.

Το παράδειγμα της εικόνας 6 είναι ενδεικτικό του τρόπου που κατασκευάζεται ο πίνακας των υπό συνθήκη πιθανοτήτων. Συγκεκριμένα ο τίτλος της αριστερής στήλης έχει πάντα την ετικέτα «Πατέρας» και ακριβώς από κάτω τοποθετούνται όλα τα ονόματα των κόμβων που επηρεάζουν άμεσα τον κόμβο της ερώτησης. Σε αυτή την περίπτωση ο κόμβος της ερώτησης έχει μόνο ένα πατέρα, με αποτέλεσμα να υπάρχει μία μόνο στήλη σε αυτή την πλευρά του πίνακα. Στη δεξιά πλευρά του πίνακα η δεξιά στήλη φέρει ετικέτα το όνομα του κόμβου με τον οποίο συσχετίζεται ο πίνακας πιθανοτήτων δηλαδή του κόμβου που ρωτάται. Ακριβώς από κάτω παρουσιάζονται οι καταστάσεις που μπορεί να έχει ο κόμβος αυτός. Στο υπόλοιπο τμήμα του πίνακα αποθηκεύονται οι πιθανότητες.



Εικόνα 6 Πίνακας υπό συνθήκη πιθανοτήτων για ένα κόμβο παιδί

Οι κόμβοι που δεν έχουν πατέρα έχουν κι αυτοί πίνακα υπό συνθήκη πιθανοτήτων. Ο πίνακας αυτός όμως περιλαμβάνει ξεχωριστά τις πιθανότητες της μεταβλητής αυτής για κάθε κατάσταση του κόμβου πατέρα (Εικόνα 7).

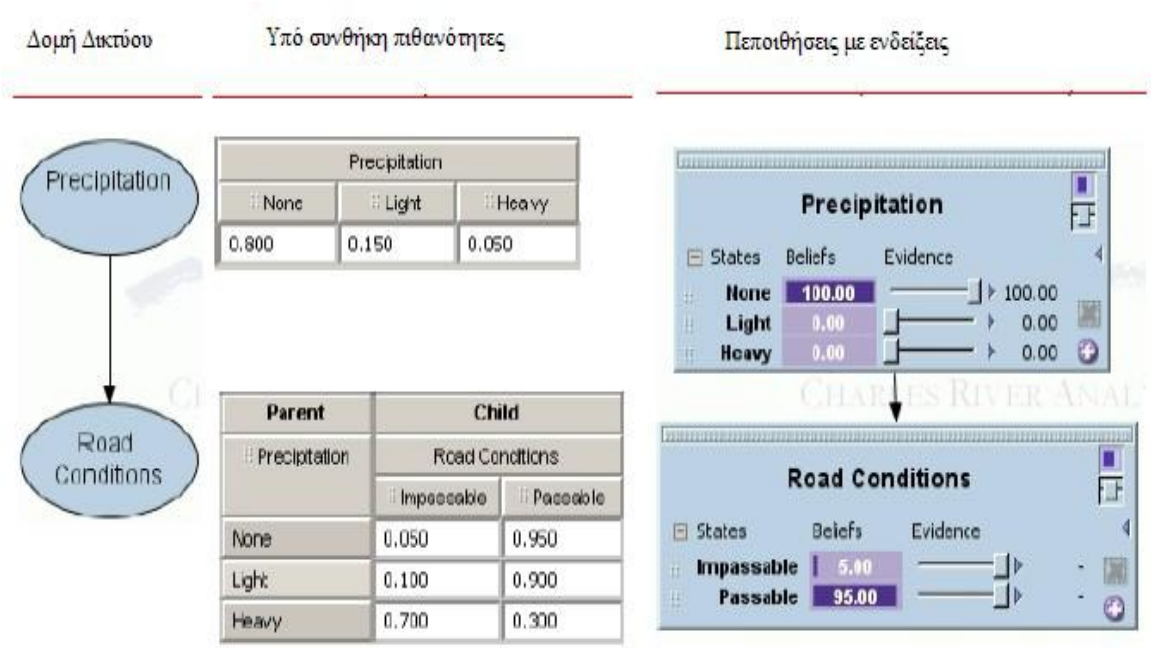


Εικόνα 7 Πίνακας υπό συνθήκη πιθανοτήτων για κόμβο-πατέρα

1.4.5 Πεποιθήσεις και Ενδείξεις

Ο όρος πεποιθήσεις (beliefs) αναφέρεται στην πιθανότητα του να είναι μια μεταβλητή σε μία συγκεκριμένη κατάσταση. Οι εκ των προτέρων πεποιθήσεις (a- priori beliefs) αποτελούν ειδική περίπτωση πεποιθήσεων που βασίζονται αποκλειστικά σε πρότερη πληροφορία. Καθορίζονται επιπλέον μόνο από την πληροφορία που έχει αποθηκευτεί στον πίνακα των υπό συνθήκη πιθανοτήτων των Bayesian δικτύων.

Η ένδειξη (evidence) είναι πληροφορία σχετικά με μία τρέχουσα κατάσταση. Για παράδειγμα στο απλό Bayesian δίκτυο της εικόνας 8, παρατηρούμε την ένδειξη ότι δεν έχει πραγματοποιηθεί πρόσφατα καμία απερισκεψία. Τα αποτελέσματα της ένδειξης στην τρέχουσα πεποίθηση απεικονίζονται στη στήλη «Πεποιθήσεις» του κόμβου «Απερισκεψία». Είμαστε πλέον απολύτως σίγουροι (100% πιθανότητα) ότι δεν υπάρχει καμία «Απερισκεψία». Τα δίκτυα Bayes υποστηρίζουν αόριστες και ατελής ενδείξεις επιτρέποντας στους χρήστες τους να εισάγουν δικές τους, πιθανότητες ένδειξης, για κάθε μεταβλητή που μπορεί να βρίσκεται στις διάφορες καταστάσεις της.



Εικόνα 8 Bayes δίκτυο, που δείχνει τα αποτελέσματα της απερισκεψία στις συνθήκες του δρόμου.

Υπάρχουν δύο διαφορετικά είδη ενδείξεων:

- Ισχυρές ενδείξεις (hard evidence): Ενδείξεις όπου ένας κόμβος βρίσκεται 100% σε μία κατάσταση και 0% σε κάθε άλλη κατάσταση (Παράδειγμα στην εικόνα 8).
- Ασθενείς ενδείξεις (soft evidence): Οποιαδήποτε άλλη ένδειξη που δεν είναι ισχυρή. Διαφορετικά ενδείξεις όπου ένας κόμβος είναι λιγότερο

του 100% σε μία κατάσταση και περισσότερο του 0% στις υπόλοιπες. Οι ασθενείς ενδείξεις χρησιμοποιούνται σε πληροφορία για την οποία υπάρχει αβεβαιότητα.

1.5 Συμπερασμός (inference) και Μάθηση (learning)

Σε πολλές περιπτώσεις, έχουμε πρόσβαση σε μεγάλες ποσότητες δεδομένων, δηλαδή σε ένα σύνολο παραδειγμάτων που δημιουργούνται από την κατανομή που θέλουμε να αναπαραστήσουμε. Αυτά τα δείγματα προφανώς μπορούν να χρησιμοποιηθούν για την κατασκευή ενός καλού μοντέλου για την υποκείμενη κατανομή, συνδυάζοντας ή όχι τη γνώση κάποιου εξειδικευμένου ανθρώπου του τομέα εφαρμογής.

1.5.1 Κατηγορίες-Στόχοι Εκμάθησης

Έστω ότι ο τομέας για τον οποίο θέλουμε να μάθουμε το μοντέλο (model learning) υπόκειται σε μία κατανομή P, η οποία προκύπτει από κάποιο είτε κατευθυνόμενο είτε μη κατευθυνόμενο μοντέλο. Θεωρούμε ένα σύνολο δεδομένων $D = \{d[1],\ldots, d[M]\}$ δειγμάτων από την P^*. Υποθέτουμε ότι τα δείγματα δειγματοληπτούνται ανεξάρτητα από την P^*, έτσι ώστε είναι ανεξάρτητα και με ίδια κατανομή (I.I.D).

Επίσης, έστω ότι έχουμε μια οικογένεια μοντέλων και ο στόχος μας είναι να μάθουμε κάποιο μοντέλο M σε αυτή την οικογένεια που να καθορίζει μια κατανομή P_M. Μπορεί να θέλουμε να μάθουμε μόνο παραμέτρους στο μοντέλο για μία σταθερή δομή, ή και να μάθουμε τη δομή του μοντέλου με βάση τα δεδομένα που μας δίνονται. (Στην περίπτωση της εκτίμησης Bayesian, αντί της εύρεσης του καλύτερου δυνατού μοντέλου ή και συνόλου παραμέτρων, μπορεί να χρειαστεί να επιστρέψουμε μια εκ των υστέρων κατανομή πάνω στις πιθανές τοπολογίες).

Συγκεκριμένα, το learning (μάθηση) μπορεί να αναφέρεται στη δομή (τοπολογία, structure) του μοντέλου ή στις παραμέτρους ή και στα δύο. Επίσης, μία σημαντική διάκριση είναι το κατά πόσον τα δεδομένα είναι πλήρως γνωστά (fully observed) ή αν κάποια από αυτά λείπουν ή δεν μπορούμε να τα παρατηρήσουμε (partially observed- hidden ή missing δεδομένα). Επομένως προκύπτουν τέσσερις δυνατοί συνδυασμοί-στόχοι του learning:

1. Γνωστή δομή, πλήρη δεδομένα: θέλουμε να υπολογίσουμε τις εκτιμήσεις μέγιστης πιθανοφάνειας (MLEs) των παραμέτρων της δεσμευμένης κατανομής
2. Γνωστή δομή, κρυμμένα ή χαμένα δεδομένα
3. Άγνωστη δομή, πλήρη δεδομένα
4. Άγνωστη δομή, μη πλήρη δεδομένα

Για τα δίκτυα Bayes, συγκεκριμένα, υπάρχουν τρεις κύριες εργασίες μάθησης-εκτίμησης:

- Εκτίμηση Παραμέτρων Μοντέλου (Parameter learning)
- Εκτίμηση μη παρατηρήσιμων μεταβλητών (Inferring unobserved variables) και
- Μάθηση Δομής (Structure learning)

1.5.2 Εκτίμηση Παραμέτρων Μοντέλου (Parameter learning)

Για να καθορίσουμε πλήρως ένα Bayesian δίκτυο και ως εκ τούτου να παρουσιάσουμε πλήρως την από κοινού κατανομή πιθανοτήτων (joint probability distribution), είναι αναγκαίο να διευκρινιστεί, για κάθε κόμβο X η κατανομή πιθανότητας για το X δοθέντων των γονιών του. Η κατανομή αυτή μπορεί να έχει οποιαδήποτε μορφή. Είναι σύνηθες να εργαζόμαστε με διακριτές ή Gaussian κατανομές δεδομένου ότι απλοποιούν τους υπολογισμούς. Μερικές φορές μόνο οι περιορισμοί σχετικά με τη διανομή είναι γνωστοί. Τότε μπορεί κανείς να χρησιμοποιήσει την αρχή της μέγιστης εντροπίας (principle of greatest entropy) για να καθορίσει μια απλή κατανομή, αυτή με τη μεγαλύτερη εντροπία, δεδομένων των περιορισμών. (Αντίστοιχα, στο ειδικό πλαίσιο ενός δυναμικού δικτύου Bayes, μπορεί κάποιος να καθορίσει, συνήθως, τη δεσμευμένη κατανομή για τη χρονική εξέλιξη της μεταβαλλόμενης κατάστασης με σκοπό τη μεγιστοποίηση του ποσοστό της εντροπίας της στοχαστικής διαδικασίας).

Συχνά, αυτές οι δεσμευμένες κατανομές περιλαμβάνουν παραμέτρους οι οποίες είναι άγνωστες και πρέπει να εκτιμηθούν από τα δεδομένα, ενίοτε κάνοντας χρήση της προσέγγισης μέγιστης πιθανοφάνειας (maximum likelihood approach). Η άμεση μεγιστοποίηση της πιθανότητας (ή της εκ των υστέρων πιθανότητας) είναι συχνά πολύπλοκη, όταν υπάρχουν μη παρατηρήσιμες μεταβλητές. Μια κλασική προσέγγιση σε αυτό το πρόβλημα είναι ο αλγόριθμος πρόβλεψης- μεγιστοποίησης (expectation- maximization algorithm), ο οποίος εναλλάσσεται μεταξύ του υπολογισμού των αναμενόμενων τιμών των μη παρατηρούμενων μεταβλητών σε εξάρτηση με δεδομένα που έχουν παρατηρηθεί και την μεγιστοποίηση της πλήρους πιθανότητας (ή μεταγενέστερης, posterior) με την προϋπόθεση ότι όλες οι τιμές που έχουν υπολογιστεί προηγουμένως είναι σωστές. Υπό κανονικές συνθήκες αυτή η διαδικασία συγκλίνει στη μέγιστη πιθανοφάνεια (ή τις μέγιστες μεταγενέστερες τιμές) για τις παραμέτρους.

Μια πιο πλήρης Bayesian προσέγγιση των παραμέτρων είναι να αντιμετωπίσουμε τις παραμέτρους σαν πρόσθετες, μη παρατηρήσιμες, μεταβλητές και να υπολογίσουμε μια πλήρη εκ των υστέρων κατανομή, πάνω σε όλους τους κόμβους κάνοντας τήρηση των δεδομένων που έχουν παρατηρηθεί και, στη συνέχεια, να ενσωματώσουν τις παραμέτρους. Αυτή η προσέγγιση μπορεί να είναι πολύπλοκη και να οδηγήσει σε μοντέλα μεγάλων διαστάσεων, έτσι στην πράξη, οι κλασσικές προσεγγίσεις καθορισμού παραμέτρων είναι πιο συχνές.

1.5.3 Συμπερασμός μη παρατηρήσιμων μεταβλητών (Inferring unobserved variables)

Επειδή ένα Bayesian δίκτυο είναι ένα ολοκληρωμένο μοντέλο για τις μεταβλητές και τις σχέσεις τους, μπορεί να χρησιμοποιηθεί για να απαντήσει σε ερωτήματα σχετικά με την πιθανότητά τους. Για παράδειγμα, το δίκτυο μπορεί να χρησιμοποιηθεί για να μαθευτεί η ενημερωμένη γνώση της κατάστασης ενός υποσυνόλου των μεταβλητών όταν παρατηρούνται άλλες μεταβλητές (οι μεταβλητές *αποδεικτικά στοιχεία*, *evidence* variables). Αυτή η διαδικασία υπολογισμού της εκ των υστέρων κατανομής των μεταβλητών βάσει αποδείξεων, ονομάζεται πιθανολογικός συμπερασμός (probabilistic inference). Η εκ των υστέρων δίνει μια καθολική επαρκή στατιστική (sufficient statistic) για εφαρμογές ανίχνευσης, όταν κάποιος θέλει να επιλέξει τιμές για ένα υποσύνολο μεταβλητών, οι οποίες ελαχιστοποιούν κάποια συνάρτηση αναμενόμενης απώλειας, όπως για παράδειγμα η πιθανότητα εσφαλμένης απόφασης (decision error). Ένα Bayesian δίκτυο, μπορεί να θεωρηθεί ένας μηχανισμός που εφαρμόζει αυτόματα το θεώρημα του Bayes σε σύνθετα προβλήματα.

Οι πιο κοινές ακριβείς μέθοδοι εκτίμησης είναι:

- **απαλοιφή μεταβλητών** (variable elimination), η οποία εξαλείφει (με συγχώνευση ή αθροιστικά) τις μη-παρατηρούμενες (non-observed) μεταβλητές, που δεν προήλθαν από αναζήτηση (non-query), μία προς μία εφαρμόζοντας την ακολουθία αθροισμάτων και μεγιστοποιήσεων του προϊόντος.
- **διάδοσης 'δέντρου της κλίκας'** (clique tree propagation), το οποίο αποθηκεύει τον υπολογισμό, έτσι ώστε πολλές μεταβλητές να μπορούν να ερωτηθούν κάθε στιγμή και τα νέα αποδεικτικά στοιχεία που μπορούν να διαδοθούν γρήγορα.
- και **αναδρομικές ρυθμίσεις** (recursive conditioning) και «**Ή / ΚΑΙ αναζήτηση**» (AND/ OR search), οι οποίες επιτρέπουν την 'εναλλαγή στον χώρο-χρόνο' (space-time tradeoff) και ισοσταθμίζουν την αποτελεσματικότητα της απαλοιφής μεταβλητών, όταν χρησιμοποιείται αρκετός χώρος.

Όλες αυτές οι μέθοδοι έχουν πολυπλοκότητα που είναι εκθετική σε σχέση με το πλάτος (treewidth) του δικτύου. Οι πιο συχνοί αλγόριθμοι προσεγγιστικής επαγωγής συμπερασμάτων (approximate inference algorithms) είναι οι : δειγματοληψίας με βάση τη σημασία (importance sampling), στοχαστική προσομοίωση Monte Carlo Markov Chain (MCMC), mini- bucket elimination, 'αναδραστική διάδοση πίστεως' (loopy belief propagation), γενικευμένη διάδοση πίστεως (generalized belief propagation), και μεταβολικές μέθοδοι (variational methods).

1.5.4 Μάθηση Δομής (Structure learning)

Στην απλούστερη περίπτωση, ένα δίκτυο Bayes καθορίζεται από έναν εμπειρογνώμονα και στη συνέχεια χρησιμοποιείται για να εκτελέσει εκτιμήσεις.

Σε άλλες εφαρμογές το έργο καθορισμού του δικτύου είναι υπερβολικά πολύπλοκο για τους ανθρώπους. Στην περίπτωση αυτή, η δομή του δικτύου και οι παράμετροι των τοπικών κατανομών θα πρέπει να γίνει γνωστή από τα δεδομένα.

Η αυτόματη εκμάθηση της γραφικής παράστασης της δομής ενός Bayesian δικτύου είναι μια πρόκληση που επιδιώκεται να επιτευχθεί στον κλάδο της 'μηχανικής μάθησης' (machine learning). Η βασική ιδέα κάνει αναδρομή σε έναν αλγόριθμο ανάκτησης που αναπτύχθηκε από τους Rebane και Pearl (1987) [3] και στηρίζεται στη διάκριση μεταξύ των τριών πιθανών τύπων παρακείμενων τριάδων (adjacent triplets) που επιτρέπονται σε ένα κατευθυνόμενο άκυκλο γράφημα (directed acyclic graph, DAG):

1. $X \rightarrow Y \rightarrow Z$
2. $X \leftarrow Y \rightarrow Z$
3. $X \rightarrow Y \leftarrow Z$

Ο τύπος 1 και ο τύπος 2 αντιπροσωπεύουν τις ίδιες εξαρτήσεις (το X και το Z είναι ανεξάρτητα δοθέντος του Y) και είναι, ως εκ τούτου, μη διακριθέντα (indistinguishable). Ο τύπος 3, ωστόσο, μπορεί να αναγνωριστεί μοναδικά, αφού τα X και Z είναι ανεξάρτητα και όλα τα άλλα ζεύγη είναι εξαρτημένα. Έτσι, ενώ οι σκελετοί (τα γραφήματα χωρίς τα βέλη) αυτών των τριών τριδύμων είναι πανομοιότυποι, η κατευθυντικότητα των βελών είναι μερικώς αναγνωρίσιμη. Η ίδια διάκριση ισχύει όταν τα X και Z έχουν κοινούς γονείς, εκτός από το ότι πρέπει κάποιος πρώτα να θέσει προϋποθέσεις για εκείνους τους γονείς. Οι αλγόριθμοι έχουν αναπτυχθεί για τον προσδιορισμό του σκελετού του υποκείμενου γραφήματος και, στη συνέχεια, όλων των βελών των οποίων η κατευθυντικότητα υπαγορεύεται από τις υπό συνθήκες ανεξαρτησίας που παρατηρήθηκαν.

Μια εναλλακτική μέθοδος εκμάθησης της διάρθρωσης χρησιμοποιεί τη διαδικασία βελτιστοποίησης με βάση την αναζήτηση (optimization based search). Απαιτεί μια μέθοδο βαθμολόγησης (scoring function) και μια στρατηγική αναζήτησης (search strategy). Μια κοινή λειτουργία βαθμολόγησης είναι η μεταγενέστερη πιθανότητα (posterior prob) της δομής δεδομένων κάποιων στοιχείων κατάρτισης (training data). Η απαίτηση του χρόνου μιας 'εξαντλητικής' -πλήρους αναζήτησης (exhaustive search), η οποία επιστρέφει τη δομή που μεγιστοποιεί τη βαθμολογία, είναι υπερεκθετική (super- exponential) του αριθμού των μεταβλητών. Μια στρατηγική τοπικής αναζήτησης κάνει σταδιακές αλλαγές που στοχεύουν στη βελτίωση της βαθμολογίας της δομής. Ένας αλγόριθμος καθολικής αναζήτησης (global search algorithm) όπως Markov Chain Monte Carlo (MCMC) μπορεί να αποφύγει το να παγιδευτεί σε τοπικά ελάχιστα (local minima). Ο Friedman και άλλοι [4] συζητήσουν τη χρησιμοποίηση αμοιβαίας πληροφόρησης μεταξύ των μεταβλητών και την εύρεση μιας δομή που να την μεγιστοποιεί αυτή. Αυτό το υλοποιούν με τον περιορισμό του συνόλου των υποψήφιων γονέων σε k κόμβους και εξαντλητική αναζήτηση σε αυτούς.

Μια άλλη μέθοδος συνίσταται στην επικέντρωση στην υποκατηγορία των μοντέλων που αποδομούνται (decomposable models), για τα οποία η εκτίμηση μέγιστης πιθανοφάνειας (MLE) έχει μια κλειστή μορφή. Είναι τότε δυνατόν να ανακαλυφθεί μια συνεπής δομή για εκατοντάδες μεταβλητών.

Ένα Bayesian δίκτυο μπορεί να αυξηθεί με κόμβους και ακμές με τη χρήση τεχνικών machine learning που βασίζεται σε κανόνες (rules- based). Επαγωγικός λογικός προγραμματισμός (inductive logic programming) μπορεί να χρησιμοποιηθεί για τους κανόνες και τη δημιουργία νέων κόμβων. Η στατιστική σχεσιακή μάθηση[5] (statistical relational learning, SRL) προσεγγίζει την χρήση μιας συνάρτησης βαθμολόγησης (score function) με βάση τη δομή του δικτύου Bayes, για να καθοδηγήσει τη δομική αναζήτηση (structural search) και να αυξήσει το δίκτυο. Μια κοινή SRL συνάρτηση βαθμολόγησης είναι η περιοχή κάτω από την καμπύλη ROC (Receiver operating characteristic).

2. ΕΚΜΑΘΗΣΗ ΔΟΜΗΣ (STRUCTURE LEARNING)

2.1 Εισαγωγή

Η πρόβλεψη δομής ή εκμάθηση της δομής αποτελεί έναν γενικό όρο για τις τεχνικές της εποπτευμένης μάθησης μηχανών (supervised machine learning) που περιλαμβάνουν την πρόβλεψη των δομών των αντικειμένων και όχι τις διακριτές τιμές ή τις πραγματικές. [6]

Για παράδειγμα, το πρόβλημα της μετάφρασης μιας πρότασης φυσικής γλώσσας (natural language) σε μια συντακτική αναπαράσταση όπως ένα συντακτικό δένδρο (parse tree) μπορεί να θεωρηθεί ως ένα πρόβλημα πρόβλεψης της δομής στο οποίο τα πεδία του αποτελέσματος είναι το σύνολο όλων των πιθανών συντακτικών δέντρων.

Τα πιθανοτικά γραφικά μοντέλα (probabilistic graphical models) αποτελούν μια μεγάλη κατηγορία μοντέλων πρόβλεψης δομής. Ειδικότερα, τα δίκτυα Bayes και τα τυχαία πεδία (random fields) είναι ευρέως χρησιμοποιούμενα για την επίλυση προβλημάτων πρόβλεψης δομής σε ένα ευρύ φάσμα τομέων εφαρμογών, συμπεριλαμβανομένης της βιοπληροφορικής, της επεξεργασία φυσικής γλώσσας, της αναγνώρισης ομιλίας (speech recognition) και όρασης υπολογιστή (computer vision). Άλλοι αλγόριθμοι και μοντέλα πρόβλεψης δομής περιλαμβάνουν τον επαγωγικό λογικό προγραμματισμό (Inductive logic programming, ILP), λογική των δικτύων Markov και υπό περιορισμούς υπό συνθήκη μοντέλα (constrained conditional models) κα.

Παρομοίως με τις τεχνικές που χρησιμοποιούνται ευρέως για εποπτευόμενη μάθηση, τα μοντέλα πρόβλεψης δομής συνήθως εκπαιδεύονται μέσω παρατηρούμενων δεδομένων, στα οποία η τιμή πραγματικής πρόβλεψης (true prediction value) χρησιμοποιείται για να ορισθούν οι παράμετροι του μοντέλου. Λόγω της πολυπλοκότητας του μοντέλου και τις αμοιβαίες σχέσεις των προβεβλημένων μεταβλητών (predicted variables), η διαδικασία της πρόβλεψης με χρήση ενός εκπαιδευμένου μοντέλου και της κατάρτισης του εαυτού του είναι συχνά, υπολογιστικά, ανέφικτο και χρησιμοποιούνται μέθοδοι προσεγγιστικής επαγωγής συμπερασμάτων (approximate inference) και μαθησιακές μέθοδοι (learning methods).

2.1.1 Έρευνα για Bayesian δίκτυα

Συγκεκριμένα, ένα Bayesian δίκτυο [7] είναι ένα πιθανοτικό γραφικό μοντέλο, που βασίζεται σε μια δομική εξάρτηση μεταξύ τυχαίων μεταβλητών, οι οποίες αντιπροσωπεύουν μια συνδυασμένη κατανομή πιθανοτήτων με ένα συμπαγή και αποδοτικό τρόπο. Αποτελείται από ένα κατευθυνόμενο, άκυκλο γράφημα (DAG), όπου οι κόμβοι απευθύνονται σε τυχαίες μεταβλητές και ορίζονται, υπό όρους, κατανομές πιθανότητας για τις μεταβλητές με βάσει τους γονείς

τους στο γράφημα. Μαθαίνοντας το γράφημα (ή την δομή) των δικτύων αυτών, από τα δεδομένα, είναι ένα από τα πιο δύσκολα προβλήματα, ακόμη και αν τα δεδομένα που έχουμε είναι πλήρη. Το πρόβλημα είναι γνωστό ως NP- hard (Chickering et al., 2003), και οι καλύτερες, ακριβείς, γνωστές μέθοδοι χρειάζονται εκθετικό χρόνο συναρτήσει του αριθμού των μεταβλητών και εφαρμόζονται σε μικρά σύνολα (περίπου 30 μεταβλητών). Οι κατά προσέγγιση διαδικασίες μπορούν να χειριστούν μεγαλύτερα δίκτυα, αλλά συνήθως έχουν προβλήματα σε τοπικά μέγιστα. Παρόλα αυτά, η ποιότητα της δομής παίζει καθοριστικό ρόλο στην ακρίβεια του μοντέλου. Αν η εξάρτηση μεταξύ των μεταβλητών δεν έχει μαθευτεί σωστά, η εκτιμώμενη κατανομή μπορεί να έχει απόκλιση από τη σωστή.

Σε γενικές γραμμές, το πρόβλημα είναι να βρούμε την καλύτερη δομή (DAG), σύμφωνα με κάποια συνάρτηση βαθμολόγησης (score function) που βασίζεται στα δεδομένα (Heckerman et al., 1995). Υπάρχουν μέθοδοι που βασίζονται σε άλλες (τοπικές) στατιστικές αναλύσεις (statistical analysis) (Spirtes et al., 1993), αλλά ακολουθούν μια εντελώς διαφορετική προσέγγιση. Η έρευνα για το θέμα αυτό είναι ανοιχτή (Chickering, 2002/ Teyssier και Koller, 2005/ Τσαμαρδίνος κ.ά., 2006/ Silander και Myllymaki, 2006/ Parviainen και Koivisto 2009/ De Campos κα, 2009/ Jaakkola κ.ά., 2010), και επικεντρωμένη κυρίως σε πλήρη δεδομένα (complete data). Σε αυτές τις περιπτώσεις, οι καλύτερες ακριβείς ιδέες (όπου είναι εγγυημένο το να βρεθεί η καλύτερη καθολικη δομή βαθμολόγησης- scoring structure) βασίζονται στον δυναμικό προγραμματισμό (Koivisto και Sood, 2004/ Singh and Moore, 2005/ Koivisto, 2006/ Silander και Myllymaki, 2006/ Parviainen και Koivisto, 2009) και χρειάζονται χρόνο και μνήμη ανάλογα με n…2n, όπου n είναι ο αριθμός των μεταβλητών. Τέτοια πολυπλοκότητα απαγορεύει τη χρήση αυτών των μεθόδων όταν υπάρχουν λίγες δεκάδες μεταβλητών, κυρίως λόγω της κατανάλωσης μνήμης (παρόλο που και η πολυπλοκότητα χρόνου είναι επίσης ένα σαφές ζήτημα). Οι Ott και Miyano (2003) επινόησαν έναν ταχύτερο αλγόριθμο όταν η πολυπλοκότητα της δομής είναι περιορισμένη (για παράδειγμα, ο μέγιστος αριθμός των γονέων ανά κόμβο και ο βαθμός συνδεσιμότητας του υποκείμενου γραφήματος). Ο Perrier κα. (2008) χρησιμοποιούν δομικούς περιορισμούς (για τη δημιουργία μιας μη κατευθυνόμενης ‘υπερδομής’ (super structure) στην οποία το μη-κατευθυνόμενο γράφημα της βέλτιστης δομής, θα πρέπει να είναι ένα υπογράφημα) για να μειώσουν το χώρο αναζήτησης. Αυτή η κατεύθυνση δείχνει να είναι ελπιδοφόρα, όταν κάποιος θέλει να μάθει τις δομές των μεγάλων συνόλων δεδομένων. Ο Kojima κα. (2010) επεκτείνουν τις ίδιες ιδέες με τη χρήση νέων στρατηγικών αναζήτησης που αξιοποιούν τις ομάδες μεταβλητών και τους περιορισμούς των προγόνων. Οι περισσότερες μέθοδοι βασίζονται στη βελτίωση της μεθόδου του δυναμικού προγραμματισμού ώστε αυτή να δουλεύει πάνω και σε μειωμένους χώρους αναζήτησης. Σε ένα διαφορετικό επίπεδο, ο Jaakkola κα. (2010) εφαρμόζουν γραμμικό προγραμματισμό χαλάρωσης (linear programming relaxation) για την επίλυση του προβλήματος, μαζί με αναζήτηση διακλάδωσης και οριοθέτησης (branch and bound search). Οι μέθοδοι διακλάδωσης και οριοθέτησης μπορούν να είναι αποτελεσματικές όταν είναι διαθέσιμα καλά όρια και περικοπές. Για παράδειγμα, αυτό συνέβη με κάποια επιτυχία στο πρόβλημα του πλανόδιου πωλητή (Traveling Salesman Problem) (Applegate κα., 2006).

2.2 Προσεγγιστική επαγωγή συμπερασμάτων (approximate inference)

2.2.1 Στατιστικός συμπερασμός (statistical inference)

Πριν από τον καθορισμό της Μπεϋζιανής συμπερασματολογίας θα πρέπει να εξετάσουμε το ευρύτερο ζήτημα του ερωτήματος ‘ποια είναι η στατιστική συμπερασματολογία;’. Πολλοί ορισμοί είναι δυνατοί, αλλά οι περισσότεροι καταλήγουν στην αρχή ότι η επαγωγική στατιστική είναι η επιστήμη της λήψης συμπερασμάτων σχετικά με τον «πληθυσμό» από ένα «δείγμα», για στοιχεία που προέρχονται από τον πληθυσμό αυτό. Αυτό από μόνο του εγείρει πολλά ερωτήματα σχετικά με το τι σημαίνει ο όρος ‘πληθυσμός’, πώς το δείγμα συνδέεται με τον πληθυσμό, πώς θα πρέπει να θεσπίσουμε τη δειγματοληψία μας, εάν όλες οι επιλογές είναι διαθέσιμες και ούτω καθεξής. Αλλά θα αφήσουμε τα ζητήματα αυτά στην άκρη, και θα επικεντρωθούμε σε ένα απλό παράδειγμα. [8]

Ας υποθέσουμε ότι η Δασική Επιτροπή επιθυμεί να εκτιμηθεί το ποσοστό των δέντρων σε ένα μεγάλο δάσος που πάσχουν από μια συγκεκριμένη ασθένεια. Είναι ανέφικτο να ελέγχει κάθε δέντρο, έτσι αποφασίζουν να επιλέξουν ένα δείγμα μόνο από n δέντρα. Ομοίως, δεν θα συζητήσουμε εδώ, πώς θα μπορούσαν να επιλέξουν το δείγμα τους, αλλά θα υποθέσουμε ότι η δειγματοληψία τους είναι τυχαία, υπό την έννοια ότι, αν θ είναι η αναλογία των δέντρων που έχουν την ασθένεια μέσα στο δάσος, τότε κάθε δέντρο στο δείγμα θα έχει την ασθένεια, ανεξάρτητα από όλα τα άλλα δέντρα στο δείγμα, με πιθανότητα θ. Συμβολίζοντας με X την τυχαία μεταβλητή που αντιστοιχεί στον αριθμό των μολυσμένων δέντρων του δείγματος, η Επιτροπή θα χρησιμοποιήσει την παρατηρηθείσα τιμή X = x, για να εξαχθεί ένα συμπέρασμα σχετικά με την παράμετρο θ του πληθυσμού. Αυτό το συμπέρασμα θα μπορούσε να λάβει τη μορφή εκτίμησης σημείου ($\hat{\theta}$ = 0.1), ενός διαστήματος εμπιστοσύνης (confidence interval) (95% σίγουριά ότι το θ βρίσκεται στο διάστημα [0.08, 0.12]), μιας δοκιμής υπόθεσης (hypothesis test) (απόρριψη της υπόθεση ότι θ < 0,07 σε ένα επίπεδο σημαντικότητας 5%) ή μιας πρόβλεψης (προβλέπει ότι το 15% των δέντρων θα έχει μολυνθεί το επόμενο έτος).

Σε κάθε περίπτωση, η γνώση της παρατηρούμενης τιμής, του δείγματος, X = x, χρησιμοποιείται για να εξάγουμε κάποια συμπεράσματα για τον χαρακτηριστικό πληθυσμό θ. Επιπλέον, αυτά τα συμπεράσματα γίνονται καθορίζοντας ένα μοντέλο πιθανοτήτων, f (x | θ), το οποίο ορίζει τον τρόπο, για μια δεδομένη τιμή του θ, με τον οποίο διανέμονται οι πιθανότητες των διαφορετικών τιμών του X.

Εδώ, για παράδειγμα, σύμφωνα με τις παραδοχές που έγιναν σχετικά με την τυχαία δειγματοληψία, το μοντέλο μας θα είναι:

$$X|\theta \sim \text{Binomial}\,(n,\theta)$$

$$f(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

Η στατιστική συμπερασματολογία, ακολούθως, ανέρχεται σε συμπέρασμα σχετικά με την παράμετρο θ του πληθυσμού με βάση την παρατήρηση X = x, και ουσιαστικά συμπεραίνουμε ότι οι τιμές του θ που δίνουν μεγάλη πιθανότητα στην τιμή του x που παρατηρήσαμε, είναι «πιο πιθανές» από εκείνες όπου αναθέτουν στο x χαμηλή πιθανότητα- αρχή της μέγιστης πιθανοφάνειας. (Σημειώνουμε ότι σε ευρύτερο πλαίσιο, η στατιστική συμπερασματολογία περιλαμβάνει επίσης τα θέματα της επιλογής μοντέλου, την επικύρωση του μοντέλου κλπ, αλλά θα περιορίσουμε την προσοχή μας στην εκτίμηση των παραμέτρων μέσα από μια παραμετρική οικογένεια μοντέλων).

Πριν προχωρήσουμε στην Bayesian συμπερασματολογία συγκεκριμένα, υπάρχουν μερικά σημεία που πρέπει να τονισθούν σχετικά με αυτή την κλασική προσεγγιστική επαγωγή σε συμπεράσματα. Το πιο βασικό σημείο είναι ότι η παράμετρος θ, ενώ δεν είναι γνωστή, αντιμετωπίζεται ως σταθερά παρά σαν τυχαία. Αυτός είναι ο ακρογωνιαίος λίθος της κλασικής θεωρίας, αλλά οδηγεί σε προβλήματα ερμηνείας. Θα θέλαμε ένα διάστημα εμπιστοσύνης 95% στο [0.08, 0.12] να σημαίνει ότι υπάρχει μια πιθανότητα 95% ότι το θ κυμαίνεται μεταξύ 0,08 και 0,12. Αυτό όμως δεν μπορεί να σημαίνει αυτό, αφού το θ δεν είναι τυχαίο: είτε είναι είτε δεν είναι στο διάστημα- η πιθανότητα δεν (και δεν μπορεί) να αναφερθεί σε αυτό. Το μόνο τυχαίο στοιχείο σε αυτό το μοντέλο πιθανοτήτων είναι τα δεδομένα, έτσι, η ορθή ερμηνεία του διαστήματος, είναι ότι αν εφαρμοστεί η διαδικασία μας ‘πολλές φορές’, τότε ‘μακροπρόθεσμα’, τα διαστήματα που κατασκευάζουμε θα περιέχουν το θ σε 95% των περιστάσεων. Όλα τα συμπεράσματα που βασίζονται στην κλασική θεωρία είναι αναγκασμένα να έχουν αυτή την μακροχρόνιας συχνότητας ερμηνεία (long- run – frequency interpretation), έστω και αν, για παράδειγμα, έχουμε μόνο ένα διάστημα, το [0.08, 0.12], για να ερμηνεύσουμε.

2.2.2 Μπεϋζιανή συμπερασματολογία (Bayesian inference)

Το γενικό πλαίσιο στο οποίο η Μπεϋζιανή συμπερασματολογία λειτουργεί είναι ταυτόσημο με το παραπάνω: υπάρχει μια παράμετρος πληθυσμού θ σχετικά με την οποία θέλουμε να εξάγουμε συμπεράσματα, και ένας μηχανισμός πιθανότητας f (x | θ), ο οποίος καθορίζει την πιθανότητα παρατήρησης διαφορετικών δεδομένων x, κάτω από διαφορετικές τιμές της παραμέτρου θ. Η θεμελιώδης διαφορά είναι, ωστόσο, ότι το θ αντιμετωπίζεται

ως μια τυχαία ποσότητα. Αυτό μπορεί να φαίνεται αρκετά ακίνδυνο, αλλά στην πραγματικότητα οδηγεί σε μια ουσιαστικά διαφορετική προσέγγιση στη στατιστική μοντελοποίηση και συμπερασματολογία.

Στην ουσία, ο συμπερασμός μας θα βασίζεται στο f (θ | x) αντί του f (x | θ). Δηλαδή στην κατανομή πιθανότητας της παραμέτρου δεδομένων των στοιχείων, παρά σε αυτήν των στοιχείων δεδομένης της παραμέτρου. Αυτό οδηγεί, με πολλούς τρόπους, σε πολύ πιο φυσικά συμπεράσματα, αλλά για να επιτευχθεί αυτό θα δούμε ότι είναι απαραίτητο να καθορίσουμε μια εκ των προτέρων κατανομή πιθανότητας, f (θ), η οποία θα αντιπροσωπεύει τις πεποιθήσεις σχετικά με την κατανομή του θ προτού να έχουμε οποιαδήποτε πληροφορία σχετικά με τα δεδομένα.

Αυτή η έννοια της εκ των προτέρων κατανομής για την παράμετρο θ βρίσκεται στο επίκεντρο του τρόπου σκέψης κατά Bayes, και ανάλογα με το αν πρόκειται για ένα συνήγορο ή έναν αντίπαλο της μεθοδολογίας, είναι είτε πλεονέκτημα κλειδί έναντι της κλασική θεωρία είτε η μεγαλύτερη παγίδα του.

Το παράδειγμα της δειγματοληψίας κέρματος.

Αυτό το παράδειγμα είναι αρκετά απλό, ώστε να απεικονίσει την Bayesian προσέγγιση για την συμπερασματολογία. Σε αυτό το παράδειγμα, η χρήση της εκ των προτέρων κατανομής είναι αναγκαία. Θα παρατηρήσουμε πώς η πιθανότητα εξακολουθεί να έχει ένα κεντρικό ρόλο στη Bayesian μέθοδο.

Πέντε νομίσματα έχουν τοποθετηθεί πάνω στο τραπέζι. Είστε αναγκασμένοι να εκτιμήσετε το ποσοστό θ αυτών των κερμάτων που είναι ‘γράμματα’, εξετάζοντας ένα δείγμα μόλις δύο νομισμάτων. Τώρα υπάρχουν μόνο 6 πιθανές τιμές του θ, γραμμένες ως M / 5 με M = 0, 1, ..., 5. Έστω X είναι ο αριθμός των ‘γραμμάτων’ στο δείγμα των δύο νομισμάτων, και ας υποθέσουμε ότι παρατηρούμε X = 1. Στη δεύτερη γραμμή του πίνακα είναι οι πιθανότητες του να παρατηρήσουμε αυτό το αποτέλεσμα, ανάλογα με την (άγνωστη) τιμή του θ. Για να βεβαιωθείτε, μπορείτε να δουλέψετε αυτές τις πιθανότητες, λαμβάνοντας για παράδειγμα το P (X = 1 | θ = 3/5).

Έχουμε 3 πλευρές ‘γράμματα’ και 2 ‘κορώνες’ στο σύνολο των 5 νομισμάτων. Βρίσκουμε 1 γράμματα και 1 κορώνα στο δείγμα των 2 νομισμάτων. Ο αριθμός των τρόπων να τα διαλέξουμε είναι: $\binom{3}{1} * \binom{2}{1}$ = 3 * 2 = 6, από ένα συνολικό αριθμό συνδυασμών $\binom{5}{2}$= 10. Η πιθανότητα είναι τότε 6/10 = 0,6.

θ	$\frac{0}{5}$	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	$\frac{5}{5}$
$f(X=1\|\theta)$	0.0	0.4	0.6	0.6	0.4	0.0
$f(\theta)$	$\frac{1}{32}$	$\frac{5}{32}$	$\frac{10}{32}$	$\frac{10}{32}$	$\frac{5}{32}$	$\frac{1}{32}$
$f(X=1\|\theta)\times f(\theta)=f(X=1,\theta)$	0	$\frac{2}{32}$	$\frac{6}{32}$	$\frac{6}{32}$	$\frac{2}{32}$	0
$f(\theta\|X=1)$	0	$\frac{4}{32}$	$\frac{12}{32}$	$\frac{12}{32}$	$\frac{4}{32}$	0

Τώρα, η δεύτερη γραμμή του πίνακα είναι η πιθανοφάνεια του θ. Δεν είναι μια κατανομή πιθανοτήτων- το άθροισμα δεν ισούται με 1. Οι πιο πιθανές τιμές του θ είναι 2/5 και 3/5, οι τιμές 0/5 και 5/5 αποκλείονται πλήρως.

2.2.3 Προσεγγιστικός συμπερασμός (approximate inference)

Σε πολλά προβλήματα εκτίμησης της δομής, η υψηλότερη βαθμολόγηση είναι δύσκολο να υπολογιστεί ακριβώς, που οδηγεί στην χρήση των κατά προσέγγιση μεθόδων συμπερασμού. Ωστόσο, όταν χρησιμοποιείται ο συμπερασμός σε έναν αλγόριθμο εκμάθησης, μια καλή προσέγγιση της βαθμολογίας μπορεί να μην είναι επαρκής. Ιδιαίτερα, η μάθηση μπορεί να αποτύχει [9], ακόμη και με μία κατά προσέγγιση μέθοδο συμπερασμού που χρησιμοποιεί αυστηρές εγγυήσεις προσέγγισης. Υπάρχουν δύο λόγοι γι 'αυτό. Πρώτον, η προσέγγιση των μεθόδων μπορεί να μειώσει αποτελεσματικά την εκφραστικότητα ενός υποκείμενου μοντέλου, καθιστώντας αδύνατη την επιλογή των παραμέτρων που δίνουν αξιόπιστα καλές προβλέψεις. Δεύτερον, οι προσεγγίσεις μπορούν να ανταποκρίνονται στις αλλαγές των παραμέτρων κατά τέτοιο τρόπο ώστε οι συνηθισμένοι αλγόριθμοι μάθησης να παραπλανηθούν. Σε αντίθεση με αυτά, δίνουμε δύο θετικά αποτελέσματα με τη μορφή ορίων μάθησης για τη χρήση λογικού προγραμματισμού χαλαρών εκτιμήσεων (LP- relaxed inference) στον δομημένο αισθητήρα (structured preceptor) και εμπειρικές ρυθμίσεις ελαχιστοποίησης του κινδύνου. Υποστηρίζουμε ότι χωρίς την κατανόηση του συνδυασμού εξαγωγής συμπερασμάτων και της μάθησης, όπως αυτούς, που είναι προσεγγιστικά συμβατοί, δεν μπορεί να είναι εγγυημένη η απόδοση της μάθησης στο πλαίσιο του προσεγγιστικού συμπερασμού.

Οι μέθοδοι προσεγγιστικής επαγωγής σε συμπέρασμα καθιστούν δυνατό το να μάθουμε ρεαλιστικά μοντέλα από μεγάλες δομές δεδομένων (big data) κάνοντας συμβιβασμούς σε υπολογιστικό χρόνο για περισσότερη ακρίβεια, όταν η ακριβής μάθηση και ο συμπερασμός είναι υπολογιστικά δυσεπίλυτα (computationally intractable).

Σημαντικές κατηγορίες μεθόδων είναι οι:

- Μεταβολικές μέθοδοι Bayesian (Variational Bayesian Methods)
- Προσδοκία διάδοσης (Expectation propagation)
- Στοχαστικά πεδία Markov (Markov random fields)

- Bayesian δίκτυα (Bayesian networks)
    - Μεταβολική μετάδοση μηνυμάτων (variational message passing)
- Με βρόγχους (loopy) και γενικευμένη (generalized) διάδοση πεποιθήσεων (belief propagation)

Όπως προαναφέραμε, τα μοντέλα πρόβλεψης δομής συνήθως περιλαμβάνουν σύνθετα προβλήματα εξαγωγής συμπερασμάτων για τα οποία το να βρεθεί η ακριβής λύση είναι δυσεπίλυτο. Υπάρχουν δύο τρόποι για την αντιμετώπιση αυτής της δυσκολίας. Άμεσα, τα μοντέλα που χρησιμοποιούνται στην προσπάθεια αυτή μπορεί να περιορίζονται σε εκείνα για τα οποία τα συμπέρασματα είναι εφικτό να βγουν, όπως είναι τα δέντρα με υπό όρους τυχαία πεδία ή τα συνειρμικά δίκτυα Markov (associative Markov networks) με δυαδικά σήματα. Γενικότερα, ωστόσο, αποτελεσματικές αλλά κατά προσέγγιση διαδικασίες εξαγωγής συμπερασμάτων έχουν επινοηθεί και εφαρμόζονται σε ένα ευρύ φάσμα των μοντέλων, συμπεριλαμβανομένων των διαδόσεων πεποιθήσεων με βρόγχους [10], μετάδοση μηνυμάτων επαναπροσδιορισμού του βάρους του δέντρου (tree reweighted message passing) [11], τα οποία παρέχουν αποτελεσματική προσέγγιση στις προβλέψεις για τα γραφικά μοντέλα ακόμα και αυθαίρετης δομής.

Από τη στιγμή που κάποια μορφή συμπερασμού είναι το κυρίαρχο υποπρόγραμμα για όλους τους αλγόριθμους μάθησης δομής, είναι φυσικό να δούμε καλές τεχνικές εξαγωγής συμπερασμάτων κατά προσέγγιση, ως λύσεις στο πρόβλημα της ‘βατής’ μάθησης (tractable learning). Ένας αριθμός συγγραφέων έχει λάβει αυτή την προσέγγιση, χρησιμοποιώντας τις κατά προσέγγιση εκτιμήσεις ως αντικατάσταση παράλειψης (drop- in replacements) κατά τη διάρκεια της εκπαίδευσης, συχνά με επιτυχία χάριν εμπειρίας. Και όμως έχει σημειωθεί ελάχιστη θεωρητική ανάλυση της σχέσης μεταξύ της κατά προσέγγισης συμπερασματολογίας και της αξιόπιστης μάθησης.

Έχουμε αποδείξει με αυτά τα παραδείγματα ότι τα χαρακτηριστικά των αλγορίθμων του κατά προσέγγιση συμπερασμού που σχετίζονται με την εκπαίδευση μπορεί να είναι διαφορετικά από άλλα, όπως οι προσεγγιστικές εγγυήσεις, που τα καθιστούν κατάλληλα για την πρόβλεψη. Πρώτον, είδαμε ότι οι προσεγγίσεις μπορούν να μειώσουν την εκφραστικότητα ενός μοντέλου, κάνοντας προηγουμένως, απλές έννοιες, αδύνατο να εφαρμοστούν και ως εκ τούτου να μάθουν, ακόμη και αν ο συμπερασμός πληροί κάποιες προσεγγιστικές εγγυήσεις. Δεύτερον, δείξαμε ότι οι τυπικοί αλγόριθμοι μάθησης μπορεί να παρασυρθούν από ανακριβή συμπερασμό, αποτυγχάνοντας να βρουν έγκυρες παραμέτρους του μοντέλου. Επομένως, είναι σημαντικό να επιλέξουμε συμβατές διαδικασίες εξαγωγής συμπερασμάτων και μάθησης.

Με αυτές τις σκέψεις στο μυαλό, έχουμε αποδείξει ότι οι μέθοδοι Lp που βασίζονται σε χαλάρωση προσεγγιστικού συμπερασμού (LP- relaxation-based approximate inference procedures) είναι συμβατές με τον αισθητήρα δομής, καθώς και με εμπειρική ελαχιστοποίηση του κινδύνου με ένα κριτήριο περιθωρίου που χρησιμοποιούν στα PAC- Bayes σύστήματα.

Σε ένα περιβάλλον πρόβλεψης, ο στόχος του κατά προσέγγιση συμπερασμού είναι να υπολογίσουμε αποτελεσματικά μια πρόβλεψη με την υψηλότερη δυνατή βαθμολογία. Ωστόσο, στην εκμάθηση, στενή σχέση μεταξύ του

μοντέλου βαθμολόγησης και της αληθινής χρησιμότητας δεν μπορεί να επιβεβαιωθεί πάντα. Εξ άλλου, η μάθηση επιδιώκει να βρει μια τέτοια σχέση.

Δεν μπορούμε να περιμένουμε να μάθουμε χωρίς αλγοριθμική διαχωριστικότητα. Κανένα μέγεθος εκπαίδευσης δεν μπορούμε να ελπίζουμε ότι θα είναι επιτυχές όταν απλά δεν υπάρχουν αποδεκτές παράμετροι του μοντέλου. Παρ όλα αυτά, θα μπορούμε να διαλέγουμε μια από τις συνήθεις τεχνικές για την αντιμετώπιση του (γεωμετρικού) μη διαχωρισμού (inseparability), σε αυτή την περίπτωση.

Ο προσεγγιστικός συμπερασμός εισάγει μια άλλη επιπλοκή, όμως. Οι τεχνικές μάθησης εκμεταλλεύονται παραδοχές σχετικά με το υποκείμενο μοντέλο για να ψάξουν το σύνολο των παραμέτρων. Ο αισθητήρας, για παράδειγμα, θεωρεί ότι η αύξηση της βαρύτητας για τα χαρακτηριστικά οδηγούν σε σωστές σημάνσεις αλλά όχι ότι λανθασμένη σήμανση θα οδηγήσει σε καλύτερες προβλέψεις. Ενώ αυτό είναι γενικά αληθές σε σχέση με ένα γραμμικό μοντέλο, ανακριβείς μέθοδοι εξαγωγής συμπερασμάτων μπορούν να διαταράξουν ή ακόμη και να αναστρέψουν τέτοιες υποθέσεις.

Μια κοινή μέθοδος προσεγγιστικού συμπερασμού είναι η διάδοση πεποιθήσεων με βρόγχους (LBP), στην οποία η διάδοση μηνυμάτων μέγιστου γινομένου (max product message passing), που είναι γνωστό ότι είναι ακριβής για τα δέντρα, εφαρμόζεται σε αυθαίρετα, κυκλικά γραφικά μοντέλα. Ενώ η LBP είναι, φυσικά, ανακριβής, η συμπεριφορά της μπορεί να είναι ακόμη πιο προβληματική για τη μάθηση. Επειδή η LBP δεν ανταποκρίνεται στις παραμέτρους του μοντέλου με τον συνήθη τρόπο, οι προβλέψεις της μπορεί να οδηγήσουν κάποιον μακριά από προσεγγιστικές παραμέτρους ακόμη και για αλγοριθμικά διαχωρίσιμα προβλήματα (algorithmically separable problems).

Συνοψίζοντας, είδαμε ότι η αποτελεσματική χρήση της κατά προσέγγιση συμπερασματολογίας για μάθηση δομής εξαρτάται από δύο παράγοντες που δεν έχουν σημασία για την πρόβλεψη. Πρώτον, η εκφραστικότητα του κατά προσέγγιση συμπερασμού, και, κατά συνέπεια, η μάθηση, μπορεί να διαφέρει σημαντικά από εκείνη του ακριβούς συμπερασμού. Δεύτερον, οι αλγόριθμοι εκμάθησης μπορεί να παρερμηνεύσουν τα δεδομένα που λαμβάνονται από τις μεθόδους προσεγγιστικού συμπερασμού, κάτι που οδηγεί σε μη επιθυμητά αποτελέσματα ή ακόμη και αποκλίσεις. Ωστόσο, όταν ισχύει σωστός διαχωρισμός αλγορίθμων, η χρήση του LP-χαλαρού συμπερασμού σε συστήματα τυπικής μάθησης δίνει αποδεδειγμένα καλά αποτελέσματα.

Η ανοιχτή έρευνα στον κλάδο περιλαμβάνει εργασίες που αφορούν έρευνα για εναλλακτικές μεθόδους συμπερασμού, οι οποίες ενώ είναι λιγότερο κατάλληλες για την πρόβλεψη και μόνο, δίνουν καλύτερη ανατροφοδότηση (feedback) για τη μάθηση. Αντίθετα, μέθοδοι εκμάθησης που είναι ειδικά προσαρμοσμένοι στις ιδιαιτερότητες συγκεκριμένων αλγορίθμων συμπερασμού, θα μπορούσαν να δείξουν βελτιωμένη απόδοση σε σχέση με εκείνους που υπονοούν ακριβή συμπερασμό. Τέλος, η έννοια της διαχωρισιμότητας αλγορίθμων και οι τρόποι με τους οποίους θα μπορούσε να σχετίζεται (μέσω της προσέγγισης) με την παραδοσιακή διαχωριστικότητα, χρήζουν περαιτέρω μελέτης.

2.3 Δυναμικός προγραμματισμός (dynamic programming)

2.3.1 Ιστορικά στοιχεία

Την μεθοδολογία του Δυναμικού Προγραμματισμού την εισήγαγε και την ανέπτυξε ο Αμερικάνος μαθηματικός Dr Richard Bellman [12] στη δεκαετία του 1950, όταν ήταν καθηγητής στο Πανεπιστήμιο της Καλιφόρνιας Berkeley και ταυτόχρονα ερευνητής στην εταιρεία Rand Corp.

Συγκεκριμένα αφού εργάστηκε πάνω σε προβλήματα αλληλοεξαρτώμενων αποφάσεων από το 1951 έως το 1957 συγκέντρωσε αρκετό υλικό ώστε να κυκλοφορήσει το πρώτο του βιβλίο για τον Δυναμικό προγραμματισμό. Στα τέλη της δεκαετίας καθιερώθηκε ως ο πρωτεργάτης της συγκεκριμένης μεθόδου, η οποία γρήγορα αναπτύχθηκε σε κεντρικό εργαλείο επίλυσης προβλημάτων Επιχειρησιακής Έρευνας.

2.3.2 Κεντρική ιδέα

Η βασική ιδέα του Δυναμικού Προγραμματισμού είναι ότι μπορούμε να χωρίσουμε κατάλληλα το πρόβλημα που αντιμετωπίζουμε σε τόσα υποπροβλήματα όσες είναι και οι άγνωστες μεταβλητές του και να προσδιορίζουμε κάθε φορά την τιμή μιας μόνο μεταβλητής. Έτσι αντί να έχουμε ένα πρόβλημα με τρείς, παραδείγματος χάρη μεταβλητές, σχηματίζουμε τρία αλληλοσυνδεόμενα υποπροβλήματα με μία μεταβλητή στο καθένα.

Ο ίδιος ο Bellman ονόμασε Δυναμικό Προγραμματισμό τον τρόπο με τον οποίο μπορούμε να λύσουμε τα προβλήματα στα οποία πρέπει να πάρουμε μια σειρά αποφάσεων που η καθεμία τους επηρεάζει τις επόμενές της και που όλες μαζί θέλουμε να δημιουργούν ένα βέλτιστο αποτέλεσμα.

2.3.3 Ορισμός

Ο πρώτος ορισμός διατυπώθηκε ως εξής: «Δυναμικός Προγραμματισμός είναι η μαθηματική θεωρία των πολυσταδιακών αποφάσεων, που αναφέρεται στη βελτιστοποίηση της λειτουργίας τους, βάσει κάποιου επιλεγμένου κριτηρίου, γνωστού ως συνάρτηση κόστους ή αντικειμενικής συνάρτησης».

Πριν προχωρήσουμε στην ανάπτυξη του ορισμού θα ήταν σκόπιμο να καθορίσουμε πληρέστερα την έννοια της διαδικασίας των πολυσταδιακών αποφάσεων. Έστω ότι έχουμε ένα φυσικό σύστημα ή διαδικασία του οποίου η κατάσταση σε μια χρονική στιγμή t καθορίζεται από ένα σύνολο τιμών των μεταβλητών του, δηλαδή από ένα διάνυσμα καταστάσεως. Οι μεταβλητές του συστήματος μπορεί να παριστάνουν διάφορα μεγέθη του, όπως συντεταγμένες θέσεως, όγκου, θερμοκρασίας πίεσης κ.τ.λ. Οι μεταβλητές αυτές μπορεί να είναι είτε αιτιοκρατικές, οπότε έχουν μια συγκεκριμένη τιμή

κάθε χρονική στιγμή, είτε τυχαίες ή πιθανολογικές, οπότε δεν έχουν συγκεκριμένη τιμή αλλά μια ορισμένη συνάρτηση κατανομής πιθανότητας των τιμών τους. Με την πάροδο του χρόνου το σύστημα υφίσταται μεταβολές είτε αυτές είναι αιτιοκρατικής μορφής είτε πιθανολογικής, και οι μεταβλητές του υπόκεινται σε μετασχηματισμούς. Σε ορισμένες από τις μεταβλητές του συστήματος που τις ονομάζουμε και ελεγχόμενες μεταβλητές, είναι δυνατό να υπαγορεύουμε εμείς τους μετασχηματισμούς που θα υποστούν σε οποιαδήποτε χρονική στιγμή.

Ο καθορισμός των μετασχηματισμών που θα επιβάλουμε στο σύστημα ισοδυναμεί με μια απόφαση. Εάν πρέπει να πάρουμε μια μόνο απόφαση για ολόκληρο το διάστημα χρόνου που εξετάζουμε, τότε έχουμε να κάνουμε με την διαδικασία απόφασης ενός μόνο σταδίου.

Εάν αντίθετα πρέπει να πάρουμε μια αλληλουχία διαδοχικών αποφάσεων, τότε και χρησιμοποιούμε τον όρο ‘πολυσταδιακή διαδικασία αποφάσεων’. Μια τέτοια αλληλουχία διαδοχικών αποφάσεων ονομάζεται ‘πολιτική’ (policy). Βέλτιστη πολιτική είναι εκείνη που βελτιστοποιεί (μεγιστοποιεί ή ελαχιστοποιεί) την τιμή της επιλεγμένης αντικειμενικής συνάρτησης της διαδικασίας.

Από τα παραπάνω γίνεται φανερό ότι τα δυναμικά προβλήματα δηλαδή εκείνα που περιλαμβάνουν στις μεταβλητές τους το χρόνο, μπορούν να αναχθούν σε πολυσταδιακές διαδικασίες αποφάσεων. Έτσι προέκυψε και η ονομασία ‘Δυναμικός’ Προγραμματισμός.

2.3.4 Αρχή βελτιστοποίησης

Για την απλοποίηση των πραγμάτων ο Bellman εξέφρασε την ‘αρχή της βελτιστοποιήσεως’ η οποία εκφράζεται ως εξής: ‘Μία βέλτιστη πολιτική έχει την ιδιότητα ότι οποιαδήποτε και αν είναι η αρχική κατάσταση της διαδικασίας και η αρχική απόφαση, οι αποφάσεις που εναπομένουν πρέπει να συνιστούν μια βέλτιστη πολιτική σε σχέση με την κατάσταση που είναι αποτέλεσμα της πρώτης απόφασης’. Έτσι, αρχίζοντας από ένα στάδιο της πολυσταδιακής διαδικασίας, για το οποίο γνωρίζουμε την αρχική κατάσταση του συστήματος, επιλύουμε τις εξισώσεις του προβλήματος για το στάδιο αυτό και ορίζουμε την κατάσταση του συστήματος κατά την αρχή του επόμενου σταδίου της διαδικασίας.

Κατόπιν χρησιμοποιώντας την κατάσταση που υπολογίσαμε σαν αρχική κατάσταση του δευτέρου σταδίου προσδιορίζουμε τη βέλτιστη απόφαση για το στάδιο αυτό και επίσης την αρχική κατάσταση του επόμενου σταδίου. Προχωρώντας κατά αυτόν τον τρόπο, μπορούμε να προσδιορίσουμε μια αλληλουχία διαδοχικών αποφάσεων που αποτελεί την βέλτιστη πολιτική για όλα τα στάδια της διαδικασίας. Η πολιτική αυτή, σύμφωνα με την αρχή βελτιστοποίησης του ‘Bellman’, αποτελεί την βέλτιστη πολιτική για ολόκληρη την πολυσταδιακή διαδικασία.

2.3.5 Η μέθοδος της αντίστροφης λύσης

Συνήθως στα προβλήματα του Δυναμικού προγραμματισμού γνωρίζουμε ή μπορούμε να καθορίσουμε την τελική κατάσταση (terminal state) της διαδικασίας που θέλουμε να βελτιστοποιήσουμε και για το λόγο αυτό η αρίθμηση των σταδίων της πολυσταδιακής διαδικασίας γίνεται κατά την αντίστροφη φορά δηλαδή από το τέλος του προβλήματος που εξετάζουμε προς την αρχή του. Αυτή είναι μια βασική χαρακτηριστική ιδιότητα των προβλημάτων δυναμικού προγραμματισμού.

2.3.6 Το μοντέλο λύσης

Από όλα τα παραπάνω συμπεραίνομε ότι η μεθοδολογία του Δυναμικού Προγραμματισμού έχει τα εξής κύρια χαρακτηριστικά:

- Χωρίζουμε το πρόβλημά μας σε τόσα στάδια όσες και οι αποφάσεις που θα πρέπει στο τέλος να ληφθούν.
- Ξεκινάμε από το τελευταίο στάδιο απόφασης και προχωρούμε προς το πρώτο αντίστροφα δηλαδή προς την αρχή του προβλήματος.
- Σε κάθε στάδιο υπολογίζουμε την καλύτερη δυνατή απόφαση, από οποιαδήποτε θέση αυτού του σταδίου, μέχρι τον τελικό προορισμό, την λύση του προβλήματος.
- Η διαδικασία αυτή θα πρέπει να είναι η πιο βέλτιστη μεταξύ όλων των πιθανών καταστάσεων αυτής της θέσης και του τελικού προορισμού. Έτσι για κάθε επόμενη θέση υπολογίζουμε την απόφαση που θα πρέπει να λάβουμε από την τωρινή μας θέση έως την επόμενη, συν την βέλτιστη απόφαση που έχουμε ήδη λάβει ως τώρα μέχρι τον προορισμό μας.

Αποτέλεσμα αυτού του μοντέλου είναι ότι όταν θα φτάσουμε στην αρχή του προβλήματος, θα αθροίσουμε όλες τις βέλτιστες αποφάσεις που έχουμε λάβει έως τώρα σε κάθε στάδιό, σε μία γενική βέλτιστη πολιτική που θα πρέπει να ακολουθήσουμε για την τελική λύση του προβλήματος.

2.3.7 Πρότυπα δυναμικού προγραμματισμού

Οι υπολογισμοί διεξάγονται σε στάδια αναλύοντας το πρόβλημα σε υποπροβλήματα, όπως αναφέραμε και παραπάνω. Κάθε υποπρόβλημα εξετάζεται ξεχωριστά με στόχο τη μείωση των υπολογισμών. Επειδή όμως τα υποπροβλήματα αλληλοσυνδέονται μεταξύ τους, πρέπει να επινοηθεί μια διαδικασία που να συνδέει τους υπολογισμούς, με τέτοιο τρόπο ώστε μια εφικτή λύση για ένα στάδιο να είναι επίσης εφικτή για το συνολικό πρόβλημα. Η μεθοδολογία του Δυναμικού προγραμματισμού θα αναπτυχθεί με το παράδειγμα του ‘ταξιδιώτη’.

Το παράδειγμα του ταξιδιώτη.

Έστω ότι ένας ταξιδιώτης θέλει να ταξιδέψει από την πόλη Α στην πόλη Λ διανύοντας την μικρότερη συνολικά χιλιομετρική απόσταση μέσα σε τέσσερις μέρες (βλέπε εικόνα 9). Μετά από μελέτη των αποστάσεων κατέληξε ότι θέλει να διανυκτερεύσει τα βράδια του, στις παρακάτω πόλεις:

Αφετηρία : πόλη

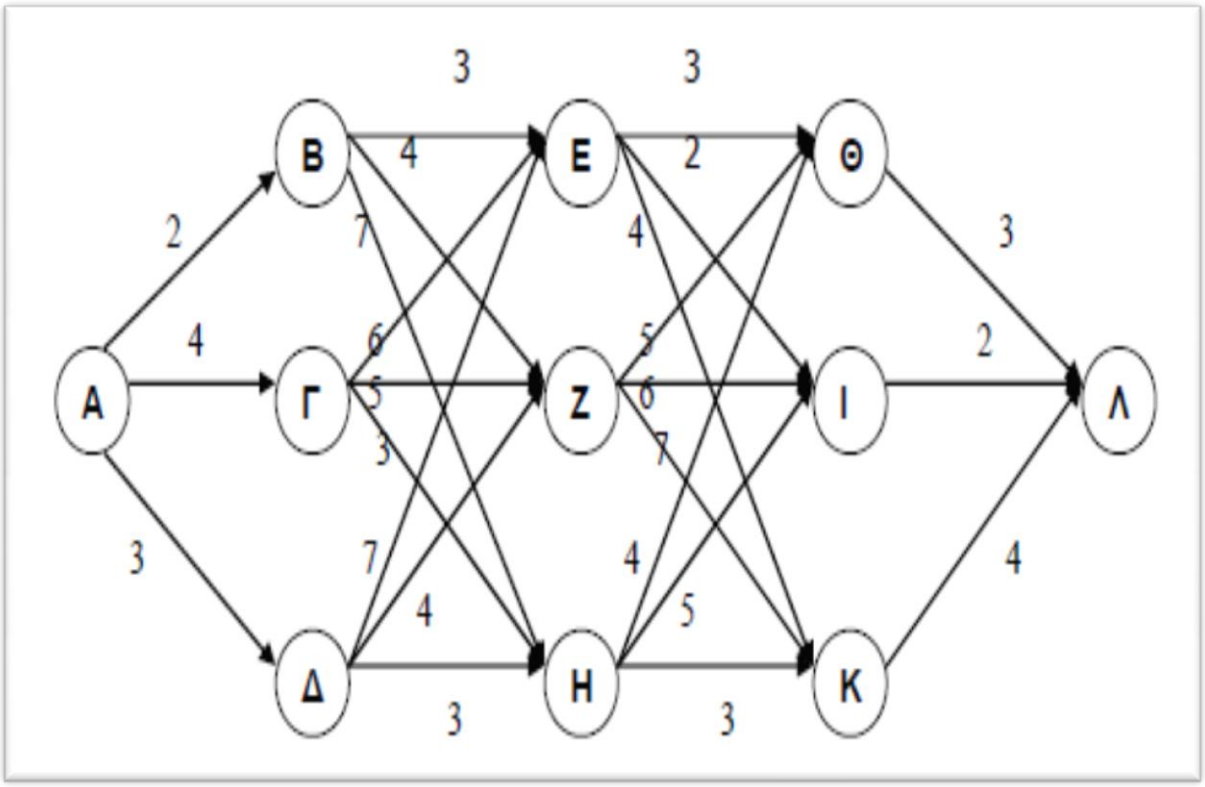
1ο βράδυ: πόλη Β, Γ ή Δ

2ο βράδυ: πόλη Ε, Ζ ή Η

3ο βράδυ: πόλη Θ, Ι ή Κ

4ο βράδυ: προορισμός στην πόλη Λ

Στην εικόνα που ακολουθεί δίνονται οι αποστάσεις μεταξύ των πόλεων, που μπορούν να καλυφθούν μέσα σε μία μέρα.



Εικόνα 9 Διάγραμμα σύντομης διαδρομής - μοντέλο "ταξιδιώτη"

Μπορούμε να παρατηρήσουμε ότι υπάρχουν (3)(3)(3)=27 διαφορετικά δρομολόγια και η μικρότερη διαδρομή βρίσκεται υπολογίζοντας την απόσταση κάθε μιας ξεχωριστά και επιλέγοντας στην συνέχεια την μικρότερη. Τέτοιου είδους προβλήματα λέγονται και συνδυαστικά (combinatorial problems). Η

μέθοδος της απαρίθμησης όλων των συνδυασμών είναι επίπονη όταν ο αριθμός των πόλεων είναι μεγάλος. Για παράδειγμα, αν υπήρχαν 10 μέρες και τρείς πιθανές πόλεις για κάθε διανυκτέρευση, ο αριθμός των συνδυασμών θα ήταν 3^9 = 19.683. Το στάδιο, στο Δυναμικό προγραμματισμό, ορίζεται σαν ένα μέρος του προβλήματος στο οποίο ανήκει ένα σύνολο αμοιβαία αποκλειστικών λύσεων και από το οποίο γίνεται επιλογή της καλύτερης λύσης. Στο προηγούμενο παράδειγμα υπάρχουν τέσσερα στάδια, με την αρίθμηση πάντα να ξεκινά από το τέλος, που αντιστοιχούν στις τέσσερις αποφάσεις για την κατεύθυνση που θα ακολουθήσει αναχωρώντας από την πόλη που βρίσκεται. Τα στάδια αλληλεξαρτώνται διότι μια συγκεκριμένη απόφαση σε κάποιο στάδιο περιορίζει τις αποφάσεις σε επόμενο στάδιο.

Η πόλη που βρίσκεται ο ταξιδιώτης όταν παίρνει μια απόφαση, ορίζει την κατάσταση του ταξιδιώτη. Έτσι, με κάθε στάδιο συσχετίζονται ορισμένες καταστάσεις. Στο παράδειγμα, το στάδιο 3 έχει τρείς καταστάσεις, την Β, την Γ ή την Δ. Σκοπός των καταστάσεων είναι η απαλοιφή της αλληλεξάρτησης που υπάρχει μεταξύ των σταδίων. Η κατάσταση παριστά το ‘σύνδεσμο’ με τα επόμενα στάδια, έτσι ώστε όταν λαμβάνεται βέλτιστη απόφαση για ένα στάδιο η απόφαση αυτή να είναι εφικτή και για το συνολικό πρόβλημα. Επιτρέπεται επιπλέον η λήψη βέλτιστων αποφάσεων για τα εναπομένοντα στάδια, χωρίς να πρέπει να γίνει έλεγχος της αποτελεσματικότητας των μελλοντικών αποφάσεων στις αποφάσεις που λήφθηκαν προηγουμένως.

Σε κάθε στάδιο και κατάσταση, η επιλογή του επόμενου σταθμού γίνεται ελαχιστοποιώντας την υπόλοιπη διαδρομή. Έχει αποδειχθεί από τον Bellman ότι εάν το κριτήριο αυτό εφαρμοστεί σε κάθε στάδιο και κατάσταση, τότε η λύση που βρίσκεται ελαχιστοποιεί την συνολική διαδρομή. ‘Αρχή βελτιστοποίησης’.

Χρησιμοποιώντας το κριτήριο αυτό στο πρόβλημά μας, προσδιορίζεται η απόφαση του ταξιδιώτη αν βρισκόταν στο στάδιο 1, στη συνέχεια στο στάδιο 2, κλπ.

Έτσι για το στάδιο 1 έχουμε:

Ελάχιστη απόσταση από το Θ στο Λ είναι 3

Ελάχιστη απόσταση από το Ι στο Λ είναι 2

Ελάχιστη απόσταση από το Κ στο Λ είναι 4

Στο στάδιο 2 η ελάχιστη απόσταση υπολογίζεται χρησιμοποιώντας τα προηγούμενα αποτελέσματα:

Ελάχιστη απόσταση από το Ε στο Λ είναι 4

(min {3+3, 2+2, 4+4} = min{6,4,8} = 4) μέσω Ι

ελάχιστη απόσταση από το Ζ στο Λ είναι 8

(min {5+3, 6+2, 7+4} = min{8,8,11} = 8) μέσω Θ ή Ι

ελάχιστη απόσταση από το Η στο Λ είναι 7

(min {4+3, 5+2, 3+4} = min{7,7,7} = 7) μέσω Θ, Ι ή Κ.

Στο στάδιο 3 η ελάχιστη απόσταση υπολογίζεται χρησιμοποιώντας τα προηγούμενα αποτελέσματα:

Ελάχιστη απόσταση από το Β στο Λ είναι 7

(min {3+4, 4+8, 7+7} = min{7,12,14} = 7) μέσω Ε

ελάχιστη απόσταση από το Γ στο Λ είναι 10

(min {6+4, 5+8, 3+7} = min{10,13,10} = 10) μέσω Ε ή Η

ελάχιστη απόσταση από το Δ στο Λ είναι 10

(min {7+4, 4+8, 3+7} = min{11,12,10} = 10) μέσω Η.

Τέλος πηγαίνοντας στο 4 στάδιο έχουμε :

ελάχιστη απόσταση από το Α στο Λ είναι 9

(min {2+7, 4+10, 3+10} = min{9,14,13} = 9) μέσω Β

χρησιμοποιώντας τους παραπάνω υπολογισμούς βρίσκεται ότι η διαδρομή ελάχιστης απόστασης είναι :

από το Α θα πρέπει να πάει στο Β

από το Β στο Ε

από τα Ε στο Ι και

από το Ι στο Λ.

2.3.8 Χαρακτηριστικά προβλημάτων δυναμικού προγραμματισμού

Το προηγούμενο παράδειγμα του προβλήματος με τον ταξιδιώτη, έχει τα χαρακτηριστικά των προβλημάτων που μπορούν να λυθούν με τη μέθοδο του Δυναμικού Προγραμματισμού. Τα χαρακτηριστικά αυτά είναι τα εξής ακόλουθα:

1. Οι αποφάσεις λαμβάνονται διαδοχικά.
2. Το πρόβλημα μπορεί να διαιρεθεί σε στάδια.
3. Κάθε στάδιο έχει ορισμένο αριθμό καταστάσεων.
4. Μια απόφαση μετασχηματίζει την παρούσα κατάσταση σε μία κατάσταση για το επόμενο στάδιο
5. Τα οφέλη συσχετίζονται με τους μετασχηματισμούς των καταστάσεων και επιδιώκεται να μεγιστοποιηθούν ή να ελαχιστοποιηθεί το κόστος.
6. Η παρούσα κατάσταση περιέχει όλες τις αναγκαίες πληροφορίες. Η διαδικασία που οδήγησε στην παρούσα κατάσταση δεν επηρεάζει τις μελλοντικές αποφάσεις
7. Κατά την εφαρμογή του κριτηρίου του Bellman:
    a) Η διαδικασία αρχίζει από το τελευταίο στάδιο απόφασης και προχωρεί προς τα επόμενα.
    b) Σε κάθε στάδιο υπολογίζεται η βέλτιστη διαδρομή από κάθε κατάσταση του σταδίου μέχρι το τελικό προορισμό. Για τον

υπολογισμό της βέλτιστης διαδρομής για μία κατάσταση υπολογίζεται για κάθε κατάσταση του επόμενου σταδίου το άθροισμα της διαδρομής από την κατάσταση που είμαστε τώρα σε μια κατάσταση επόμενου σταδίου, και σε αυτήν προστίθεται η βέλτιστη υπόλοιπη διαδρομή από την κατάσταση του επόμενου σταδίου μέχρι τον τελικό προορισμό. Στη συνέχεια επιλέγεται το βέλτιστο άθροισμα και έτσι υπολογίζεται η βέλτιστη απόσταση μεταξύ της κατάστασης που βρισκόμαστε και του τελικού προορισμού.

2.3.9 Δυναμικός Προγραμματισμός σε δίκτυα Bayes

Η εκμάθηση της δομής ενός Bayesian δικτύου, από τα δεδομένα, είναι μια διαδικασία που έχει κίνητρα, αλλά είναι ταυτόχρονα και υπολογιστικά δύσκολη. Στην μελέτη των Koivisto, Sood [13] παρουσιάζεται ένας αλγόριθμος που υπολογίζει την ακριβή εκ των προτέρων πιθανότητα ενός υποδικτύου, π.χ., μια κατευθυνόμενη ακμή (directed edge). Μια τροποποιημένη εκδοχή του αλγορίθμου βρίσκει μία από τις πιο πιθανές δομές δικτύου. Αυτός ο αλγόριθμος τρέχει σε χρόνο O ($n2^n + n^{k+1} C (m)$), όπου n είναι ο αριθμός των μεταβλητών δικτύου, k είναι μια σταθερά στο μέγιστο βαθμό και C (m) είναι το κόστος υπολογισμού μιας ενιαίας, τοπικής, οριακής, υπό συνθήκη πιθανότητας για στιγμιότυπα m δεδομένων. Αυτός είναι ο πρώτος αλγόριθμος με μικρότερη από υπερ- εκθετική πολυπλοκότητα σε σχέση με το n. Ακριβής υπολογισμός μας επιτρέπει να αντιμετωπίζουμε πολύπλοκες περιπτώσεις όπου οι υπάρχουσες μέθοδοι Monte Carlo και οι διαδικασίες τοπικής αναζήτησης (local search) συνήθως αποτυγχάνουν. Δείχνεται επίσης, ότι και σε τομείς με μεγάλο αριθμό μεταβλητών, ο ακριβής υπολογισμός είναι εφικτός, δεδομένων κατάλληλων a priori περιορισμών πάνω στις δομές. Ο συνδυασμός από ακριβείς και ανακριβείς μεθόδους είναι επίσης δυνατός. Οι Koivisto και Sood παρουσιάζουν την εφαρμοσιμότητα του αλγορίθμου τους σε τέσσερα συνθετικά σύνολα δεδομένων των 17, 22, 37, και 100 μεταβλητών.

Η ανακάλυψη της δομής σε δίκτυα Bayes είχε προσελκύσει μια μεγάλη έρευνα κατά την προηγούμενη δεκαετία. Ένα Bayesian δίκτυο καθορίζει μια συνδυασμένη κατανομή πιθανότητας σε ένα σύνολο τυχαίων μεταβλητών με έναν δομημένο τρόπο. Ένα βασικό συστατικό σε αυτό το μοντέλο είναι η δομή του δικτύου, ένα κατευθυνόμενο άκυκλο γράφημα σχετικά με τις μεταβλητές, το οποίο κωδικοποιεί μια σειρά από υπό όρους ισχυρισμούς ανεξαρτησίας. Η μάθηση των άγνωστων εξαρτήσεων των στοιχείων, υποκινείται από μια ευρεία συλλογή εφαρμογών της πρόβλεψης και συμπερασμού (Heckerman et al., 1995b).

Μέθοδοι κατά Bayes, για τη μάθηση της δομής, αφορούν την εκ των υστέρων κατανομή των δομών του δικτύου. Διαφορετικές εφαρμογές απαιτούν και διαφορετικούς τύπους εκ των υστέρων αθροισμάτων (posterior summaries), οι οποίες είναι συνήθως δύσκολο να υπολογιστούν. Για παράδειγμα, όταν το ενδιαφέρον είναι καθαρά στην πρόβλεψη μελλοντικών παρατηρήσεων, τότε αυτές θα πρέπει να ενσωματωθούν κατά τη διάρκεια της εκ των υστέρων κατανομής με τον τρόπο του μέσου όρου μοντέλου (model averaging). Όταν το ενδιαφέρον είναι καθαρά επαγωγικό, τότε μπορεί κανείς να ψάξει για τις

δομές, ή τα τοπικά δομικά χαρακτηριστικά, τα οποία είναι πιο πιθανά. Από τις δημιουργικές εργασίες των Buntine (1991) και Cooper και Herskovits (1992) πολλοί αλγόριθμοι έχουν παρουσιαστεί για αυτές τις μεθόδους μάθησης της δομής. Ωστόσο, επειδή ο αριθμός των πιθανών δομών αναπτύσσεται υπερ-εκθετικά σε σχέση με τον αριθμό των μεταβλητών του δικτύου, ακριβής υπολογισμοί θεωρούνται συχνά ανέφικτοι. Πράγματι, είναι γνωστό ότι η εξεύρεση μιας βέλτιστης Bayesian δομής του δικτύου είναι NP- δύσκολο ακόμη και όταν η μέγιστη τιμή των εισερχομένων (maximum in- degree) οριοθετείται από μια σταθερά μεγαλύτερη του ενός (Chickering et al., 1995). Κατά συνέπεια, μεγάλο μέρος της έρευνας έχει επικεντρωθεί σε ανακριβείς μεθόδους (inexact methods).

Για να βρεθεί μια καλή ή μια βέλτιστη δομή του δικτύου, διάφοροι γενικοί ευριστικοί αλγόριθμοι, όπως στοχαστική τοπική αναζήτηση και γενετικοί αλγόριθμοι, έχουν χρησιμοποιηθεί (βλέπε, πχ Heckerman κά, 1995, Larranaga κά, 1996). Αυτές οι μέθοδοι μπορεί να επεκταθούν για να βρεθούν κλάσεις ισοδυναμίας των δομών του δικτύου (βλέπε Chickering, 2002, Acid και de Campos, 2003, Castelo και Kocka, 2003). Ένα κεντρικό πρόβλημα σε όλους αυτούς τους αλγόριθμους, είναι ότι κανείς δεν μπορεί να εγγυηθεί την ποιότητα του αποτελέσματος. Επίσης, η απαίτηση σε χρόνο, αν και συχνά είναι πρακτικό, μπορεί να είναι δύσκολο να εκτιμηθεί εκ των προτέρων.

Η ανακάλυψη πολύ πιθανών δομικών χαρακτηριστικών δεν έχει μελετηθεί τόσο εκτεταμένα. Οι Madigan και York (1995) προτείνουν μια μέθοδο Markov Chain Monte Carlo (MCMC) στο χώρο των δομών του δικτύου. Οι Friedman και Koller (2003) σχεδιάσανε μια πιο αποτελεσματική διαδικασία MCMC στο χώρο των τάξεων των μεταβλητών. Αυτά τα αποτελέσματα των αλγορίθμων προσεγγίζουν την εκ των υστέρων πιθανότητα των δομικών χαρακτηριστικών. Η ποιότητα της προσέγγισης δεν είναι εγγυημένη σε πεπερασμένο αριθμό εκτελέσεων.

Οι ακριβείς αλγόριθμοι για την εκμάθηση δομής έχουν παρουσιαστεί για πολύ περιορισμένες τάξεις των Bayesian δικτύων μόνο. Ο αλγόριθμος από τους Cooper και Herskovits (1992) είναι πολυώνυμο του αριθμού των μεταβλητών, αλλά μια συνεπής διάταξη των μεταβλητών θεωρείται στην είσοδο. Οι Chow και Liu (1968) δίνουν έναν αποδοτικό αλγόριθμο για την εκμάθηση δομών δέντρων (δηλαδή, κατ 'ανώτατο όριο σε βαθμό είναι το ένα). Ωστόσο, οι πιο αποδοτικοί ακριβείς αλγόριθμοι, μέχρι στιγμής, που επιτρέπουν αυθαίρετες κατασκευές δικτύου έχουν υπερ- εκθετική χρονική πολυπλοκότητα (Cooper και Herskovits, 1992, Friedman και Koller, 2003). Κατά συνέπεια, μπορούν να εφαρμοστούν μόνο όταν ο αριθμός των μεταβλητών είναι πολύ μικρός (δηλαδή, το πολύ 10). Είναι ενδιαφέρον, ωστόσο, αυτό που ο Chickering (2002) δείχνει, δηλαδή ότι υπό ορισμένες παραδοχές μονοτονίας, μια μέθοδος άπληστης αναζήτησης (greedy search method) θα βρει μια, λεγόμενη ‘βέλτιστη ένταξη δομής του δίκτυο’ (inclusion optimal network structure) όταν το μέγεθος των δεδομένων τείνει στο άπειρο, αλλά η απαίτηση του χρόνου μπορεί να είναι εκθετική.

Στην εργασία των Koivisto και Sood, προτείνεται ένας νέος ακριβής αλγόριθμος για την εύρεση της δομής σε δίκτυα Bayes ενός μέτριου μεγέθους (ας πούμε, 25 μεταβλητές ή λιγότερο). Στην πραγματικότητα, θεωρούνται δύο εκδοχές του αλγορίθμου:

1. μία για τον υπολογισμό των εκ των υστέρων πιθανοτήτων των δομικών χαρακτηριστικών
2. μια άλλη για την εξεύρεση της βέλτιστης δομής.

Κάνουν επίσης μια αυστηρή ανάλυση πολυπλοκότητας του αλγορίθμου. Το έργο αυτό υπαγορεύεται από τρεις σειριακές παρατηρήσεις, οι οποίες συνοψίζονται παρακάτω.

Πρώτον, μπορούμε να αναμένουμε ότι οι υπάρχουσες μέθοδοι αποτυγχάνουν στις περιπτώσεις όπου το εκ των υστέρων τοπίο (posterior landscape) των δομών του δικτύου είναι ιδιαίτερα περίπλοκο, που συμπεριλαμβάνει και πολλαπλούς ‘τρόπους’ (multiple ‘modes’). Αυτή η περίπτωση είναι ιδιαίτερα εμφανής, όταν οι υποκείμενες εξαρτήσεις δεν δείχνουν κανένα σήμα του περιθωρίου (no marginal signals). Δηλαδή, οι ακμές δεν μπορεί να ανιχνευθούν με βάση κατά ζεύγη συσχετισμούς (οι οποίοι θα μπορούσαν να είναι και όλοι μηδενικοί). Στη συνέχεια, είναι αναγκαίο να εξετασθούν όλες οι πιθανές τοπικές δομικές εξαρτήσεις. Αυτό παρακινεί τις προσπάθειες για την επέκταση του πεδίου εφαρμογής των ακριβείς υπολογισμών.

Δεύτερον, αποδεικνύεται ότι υπάρχουν ουσιαστικά δύο παράλληλες συνεισφορές στην πολυπλοκότητα του ακριβούς υπολογισμού. Η μια οφείλεται στην εξέταση όλων των πιθανών τοπικών δομικών εξαρτήσεων υπό την οπτική γωνία των δεδομένων. Αυτό το μέρος είναι αναπόφευκτο και μπορεί, στην πράξη, πραγματικά να κυριαρχεί στη συνολική πολυπλοκότητα. Επιπρόσθετα σε αυτό, μια άλλη συμβολή οφείλεται στην εξερεύνηση όλων των γραφημάτων δομών. Αν και αυτό φαίνεται να διαρκεί περισσότερο με πολυωνυμική καθυστέρηση, αποδεικνύεται ότι το μέρος αυτό δεν εξαρτάται από το μέγεθος των δεδομένων ούτε την πολυπλοκότητα των τοπικών μοντέλων. Έτσι, για ένα αρκετά μικρό αριθμό μεταβλητών δικτύου, στην πράξη μπορούμε να αντέξουμε ‘οικονομικά’ την ακριβή αυτή διερεύνηση όλων των δομών του δικτύου.

Η τρίτη παρατήρηση είναι ότι οι υπάρχοντες ακριβείς αλγόριθμοι (Cooper και Herskovits, 1992, Friedman και Koller, 2003) μπορούν να βελτιωθούν σημαντικά. Πιο συγκεκριμένα, παρουσιάζεται στον πρώτο αλγόριθμο τους η εύρεση μέσων όρων (ή, εναλλακτικά, μεγιστοποιήσεων) πάνω σε όλες τις δομές του δικτύου σε λιγότερο από υπερ- εκθετικό χρόνο.

3. ΕΦΑΡΜΟΓΗ (IMPLEMENTATION)

3.1 Aition Desk

3.1.1 Γενικά

Σε αυτό το σημεία, παρέχεται μια επισκόπηση της αρχιτεκτονικής του συστήματος AITION [14]. Ο στόχος του παρόντος υποκεφαλαίου είναι να δώσει μια σύντομη κατανόηση των επιχειρησιακών και τεχνικών προδιαγραφών που καλύπτονται από το σύστημα και μια συνολική περιγραφή των δομικών μονάδων της εφαρμογής (application modules), τόσο από αλγοριθμική όσο και από χρηστική άποψη.

3.1.2 Επισκόπηση του συστήματος

Σε αυτό το υποκεφάλαιο λοιπόν, θα παρουσιάσουμε το ‘AITION’, ένα ολοκληρωμένο (integrated) αιτιολογικό (causal) πιθανοτικό (probabilistic) σύστημα δικτύου (network framework) με γνώμονα την οντολογία (ontology-driven) για ανακάλυψη γνώσης, επιλογή χαρακτηριστικών, κάθετη ολοκλήρωση (vertical integration) και σημασιολογική μοντελοποίηση της υπό συνθήκη αβεβαιότητας.

Το σύστημα έχει αναπτυχθεί από την ομάδα MaDgIK Lab Team του ΕΚΠΑ με σκοπό να βοηθήσει γιατρούς να αναλύσουν ετερογενή δεδομένα και να βρουν συσχετίσεις και εξαρτήσεις μεταξύ διαφορετικών ιατρικών μεταβλητών

3.1.3 Εισαγωγή στο AITION

Στην εποχή μας υπάρχει μια μεγάλη ανάγκη για την συγχώνευση των πληροφοριών στους διάφορους αφαιρετικούς τομείς ενός συστήματος, σύμφωνα με την οποία όλα τα στρώματα συστήματος αυτού είναι καθέτως ολοκληρωμένα, προκειμένου να υπάρχει μια ενιαία εικόνα της κατάστασης του.

Για τον έλεγχο των συνεχώς αυξανόμενης πολυπλοκότητας συστημάτων, είναι σημαντικό να έχουμε μια σωστή κατανόηση των διεργασιών τους. Τα χαρακτηριστικά των διεργασιών μπορεί να ποικίλουν πολύ και επιπλέον είναι αβέβαια. Αρκετές, κρίσιμες προκλήσεις πρέπει να αντιμετωπιστούν, όπως η εννοιολογική ετερογένεια (conceptual heterogeneity) (επιπλέον στις πιο συνήθεις: σημασιολογική (semantic) και συντακτική (syntactic)), καθώς και η ελλιπής πληροφόρηση και η αβεβαιότητα. Η αβεβαιότητα καθιστά ιδιαίτερα δύσκολο, τη συνολική κατανόηση να επιτευχθεί.

Επιπλέον, η αναγνώριση των σχετικών μεταβλητών (που ονομάζονται επίσης χαρακτηριστικά (features)) είναι ένα βασικό συστατικό της ανακάλυψης γνώσης, της ποσοτικής μοντελοποίησης, της κατασκευής μοντέλων υποστήριξης αποφάσεων (decision support models) οδηγούμενα από δεδομένα, και της ανακάλυψης με τη βοήθεια υπολογιστή (computer- assisted discovery). Οι περισσότερες μέθοδοι επιλογής χαρακτηριστικών δεν επιχειρούν να αποκαλύψουν αιτιώδεις σχέσεις μεταξύ των χαρακτηριστικών και στόχων, και αντιθέτως επικεντρώνονται στην εξαγωγή καλύτερων προβλέψεων

Τέλος, η σημασιολογική διαλειτουργικότητα (semantic interoperability) είναι απαραίτητη για την διαλειτουργικότητα, την ανταλλαγή γνώσεων (knowledge sharing) και την επαναχρησιμοποίηση (reuse) της γνώσης, καθώς έχει την ικανότητα να υποστηρίξει την αβεβαιότητα. Καθώς μεγαλώνει η εκτίμηση των περιορισμών της οντολογικής τοπολογίας (ontology formalisms) που δεν μπορεί να συμπεριλάβει την αβεβαιότητα, αυξάνεται η ζήτηση, από τις κοινότητες χρηστών, για οντολογικές τοπολογίες με τη δυνατότητα να εκφράζουν την αβεβαιότητα.

Προκειμένου να αντιμετωπίσουμε όλα τα παραπάνω ζητήματα, υπάρχει η δυνατότητα χρήσης του AITION, που όπως προαναφέραμε είναι ένα ολοκληρωμένο αιτιολογικό πιθανοτικό σύστημα δικτύου με γνώμονα την οντολογία για ανακάλυψη γνώσης, επιλογή χαρακτηριστικών, κάθετης ολοκλήρωση και σημασιολογικής μοντελοποίησης υπό συνθήκες αβεβαιότητας.

Το AITION παρέχει τις απαραίτητες τεχνικές, αλγόριθμους, τοπολογίες (formalisms) και εργαλεία για:

- Αιτιολογική ανακάλυψη και κάθετη ολοκλήρωση, ακόμα και στην απουσία πειραματισμού που βασίζεται σε παρατηρούμενα δεδομένα, περιορισμών οντολογίας και εκ των προτέρων γνώσης.
- Επιλογή χαρακτηριστικών με βάση τη γνώση αιτιακών σχέσεων (causal relationships).
- Σημασιολογική μοντελοποίηση και διαλειτουργικότητα υπό συνθήκες αβεβαιότητας, επεκτείνοντας οντολογικές τοπολογίες με τη δυνατότητα να εκφράζουν αβεβαιότητα.

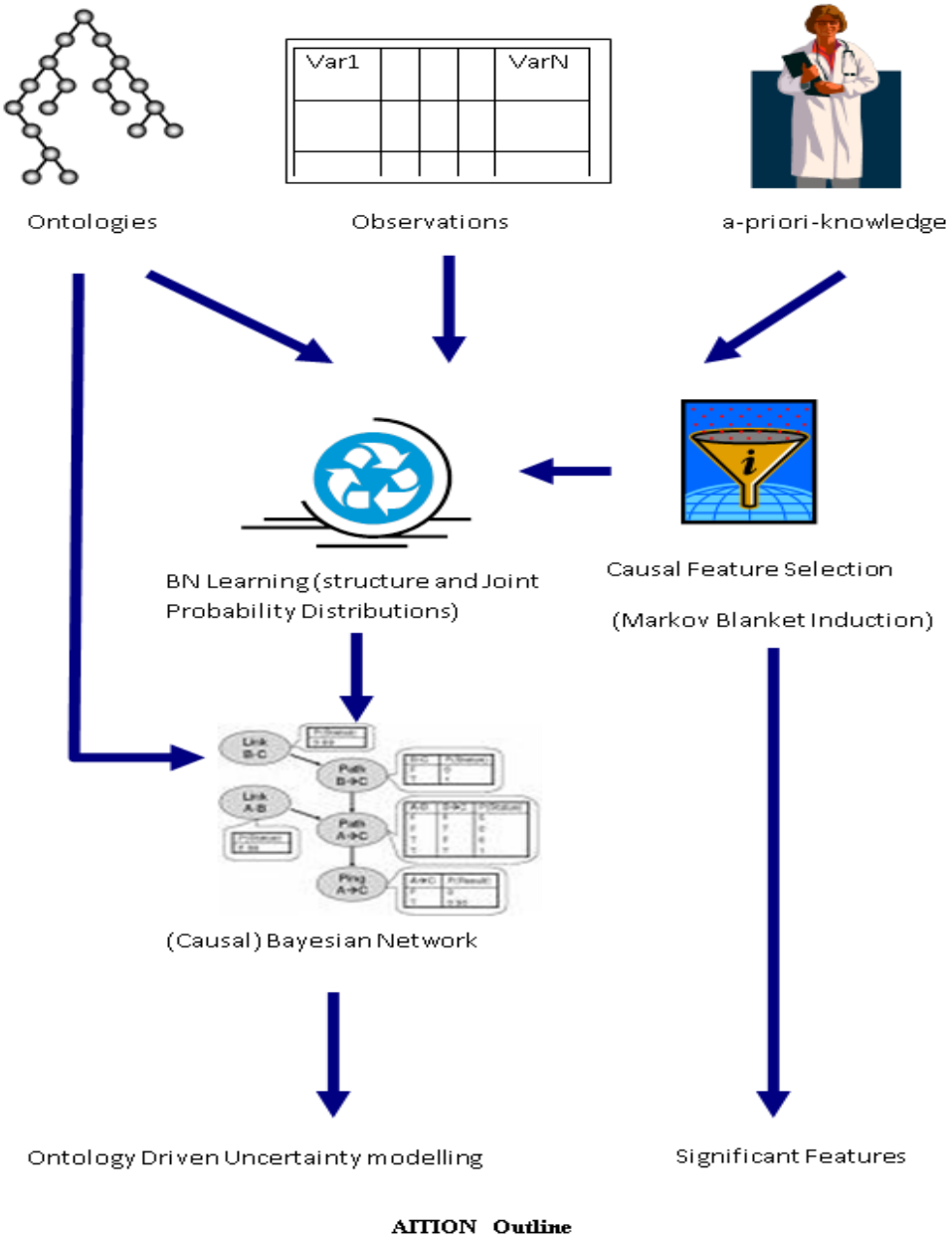
3.1.4 Βασικά χαρακτηριστικά του συστήματος

Ο σκοπός του AITION είναι να παρέχει τα απαραίτητα εργαλεία για την αιτιώδη εξερεύνηση στην περίπτωση απουσίας πειραματισμού. Άριστοι αλγόριθμοι και τεχνικές έχουν εφαρμοστεί για μάθηση δικτύων Bayes και επαγωγή σε Markov κάλυμμα (Markov blanket induction) με στόχο την αιτιώδη εξερεύνηση και επιλογή χαρακτηριστικών για καθέτως ολοκληρωμένα, ετερογενή δεδομένα. Επιπλέον, οι οντολογίες ενσωματώνονται με τα δίκτυα Bayes για την αυτοματοποίηση αιτιώδους ανακάλυψης και αιτιώδους επιλογής χαρακτηριστικών και σημασιολογική μοντελοποίηση υπό συνθήκες αβεβαιότητας.

Συνοψίζοντας, το AITION προτείνει ένα ολοκληρωμένο πλαίσιο με τα ακόλουθα χαρακτηριστικά:

- Αιτιώδης ανακάλυψη και κάθετη ολοκλήρωση βασισμένη στην αιτιώδη πιθανοτική μάθηση Δικτύων (Bayesian Δίκτυα).
- Αιτιώδης επιλογή χαρακτηριστικών χρησιμοποιώντας Markov Blanket επαγωγή και αιτιώδους ανακάλυψης της τοπικής δομής.
- Μοντέλα αβεβαιότητας οδηγούμενα από Οντολογίες, ενσωματώνοντας Bayesian Δίκτυα στη γλώσσα οντολογικού ιστού (Web Ontology Language, OWL).

Το διάγραμμα που ακολουθεί σκιαγραφεί το προτεινόμενο πλαίσιο:

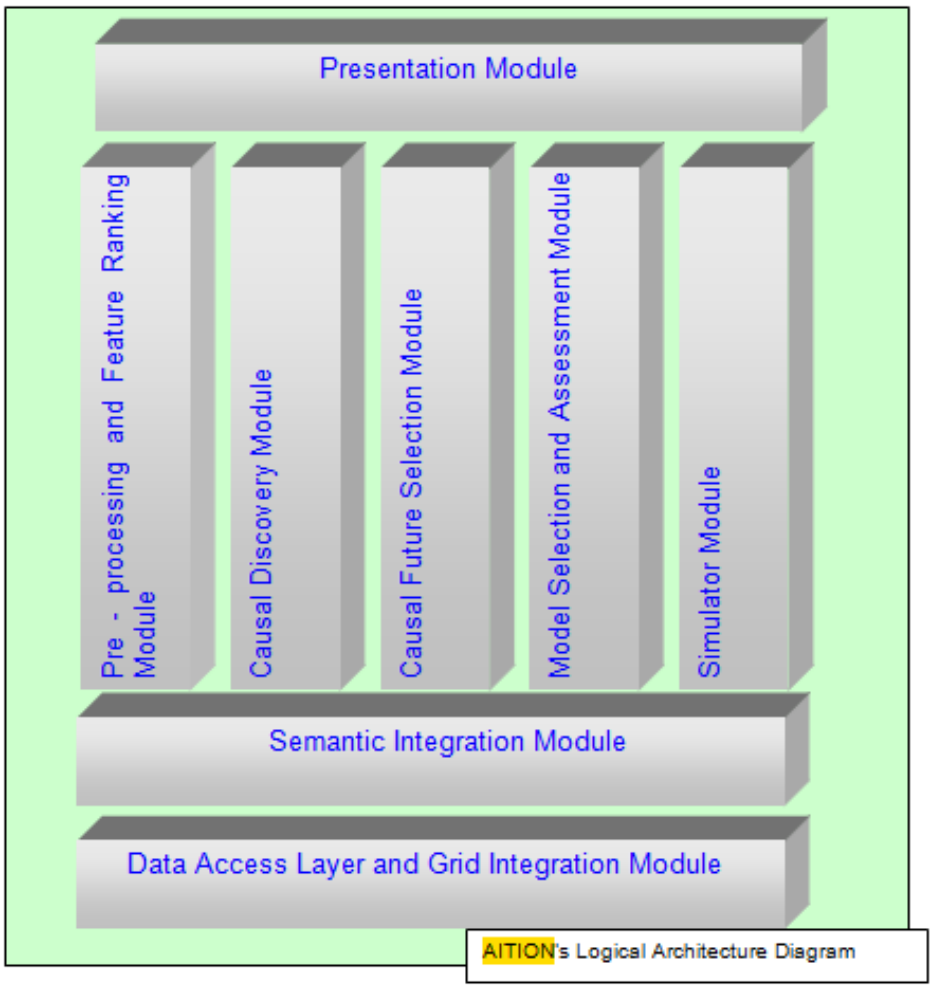


Εικόνα 10 Η εικόνα του AITION

3.1.5 Λογική Αρχιτεκτονική και Ενότητες

Το AITION αποτελείται από τις ακόλουθες ενότητες (modules):

- Προ- επεξεργασία (Pre processing) και Ενότητα Κατάταξης Χαρακτηριστικών (Feature ranking module).
- Ενότητα Αιτιώδους ανακάλυψης (Causal Discovery Module)
- Ενότητα Αιτιώδους Επιλογής Μέλλοντος (Causal Future Selection Module)
- Επιλογή Μοντέλου και Ενότητα Αξιολόγησης (Model Selection and Assessment Module)
- Μονάδα Προσομοιωτή (Simulator Module)
- Ενότητα Σημασιολογικής Ολοκλήρωσης (Semantic Integration Module)
- Ενότητα Παρουσίασης (Presentation Module)
- Επίπεδο Πρόσβασης Δεδομένων (Data Access Layer) και Ενότητα Πλέγματος Ολοκλήρωσης (Grid Integration Module)



Εικόνα 11 Λογικό διάγραμμα αρχιτεκτονικής του AITION

3.1.6 Οφέλη

- Αιτιώδης Ανακάλυψη και Κατανόηση των δεδομένων:
  - ορθή κατανόηση των διαδικασιών, μαζί με την ικανότητα να αιτιολογεί για αυτές.
  - Απόκτηση ενός μοντέλου του υποκείμενου μηχανισμού παραγωγής δεδομένων και ανάθεση πιθανοτήτων στις αιτιακές εξηγήσεις (causal explanations).
- Αιτιώδης Δυνατότητα επιλογής:
  - εξήγηση συνάφειας και σημασίας από την άποψη αιτιακών μηχανισμών,
  - διάκριση μεταξύ των πραγματικών χαρακτηριστικών και πειραματικών κατασκευασμάτων,
  - αντιμετώπιση δυσκολιών που αφορούν διαστάσεις (dimensions).
- Μοντέλα αβεβαιότητας οδηγούμενα από οντολογίες:
  - περιγραφή της γνώσεις σχετικά με έναν τομέα με την καθορισμένη αβεβαιότητα του, με έναν ηθικό, δομημένο, διαμοιραζόμενο και κατανοητό- από- μηχανές τρόπο.
  - Σημασιολογία και μοντελοποίηση δεδομένων που βασίζεται σε ολοκληρωμένα, μη ελλιπή, ετερογενή και αβέβαια κάθετα δεδομένα (vertical data).
- Αιτιολόγηση υπό συνθήκες αβεβαιότητας.

3.1.7 Τωρινή κατάσταση

Για το σκοπό αυτό, έχει αναπτυχθεί αυτή η βιβλιοθήκη λογισμικού που υλοποιεί μια σουίτα αλγόριθμων επαγωγής Αιτιώδους Πιθανοτικών δικτύων που κλιμακώνονται από χιλιάδες μεταβλητές και ως εκ τούτου είναι ιδιαίτερα κατάλληλο για την μοντελοποίηση μαζικών, μεγάλων διαστάσεων, συνόλων δεδομένων ετερογενών δεδομένων. Η εργαλειοθήκη (toolkit) εστιάζει στους αλγόριθμους αιτιώδους ανακάλυψης και επιλογής χαρακτηριστικών.

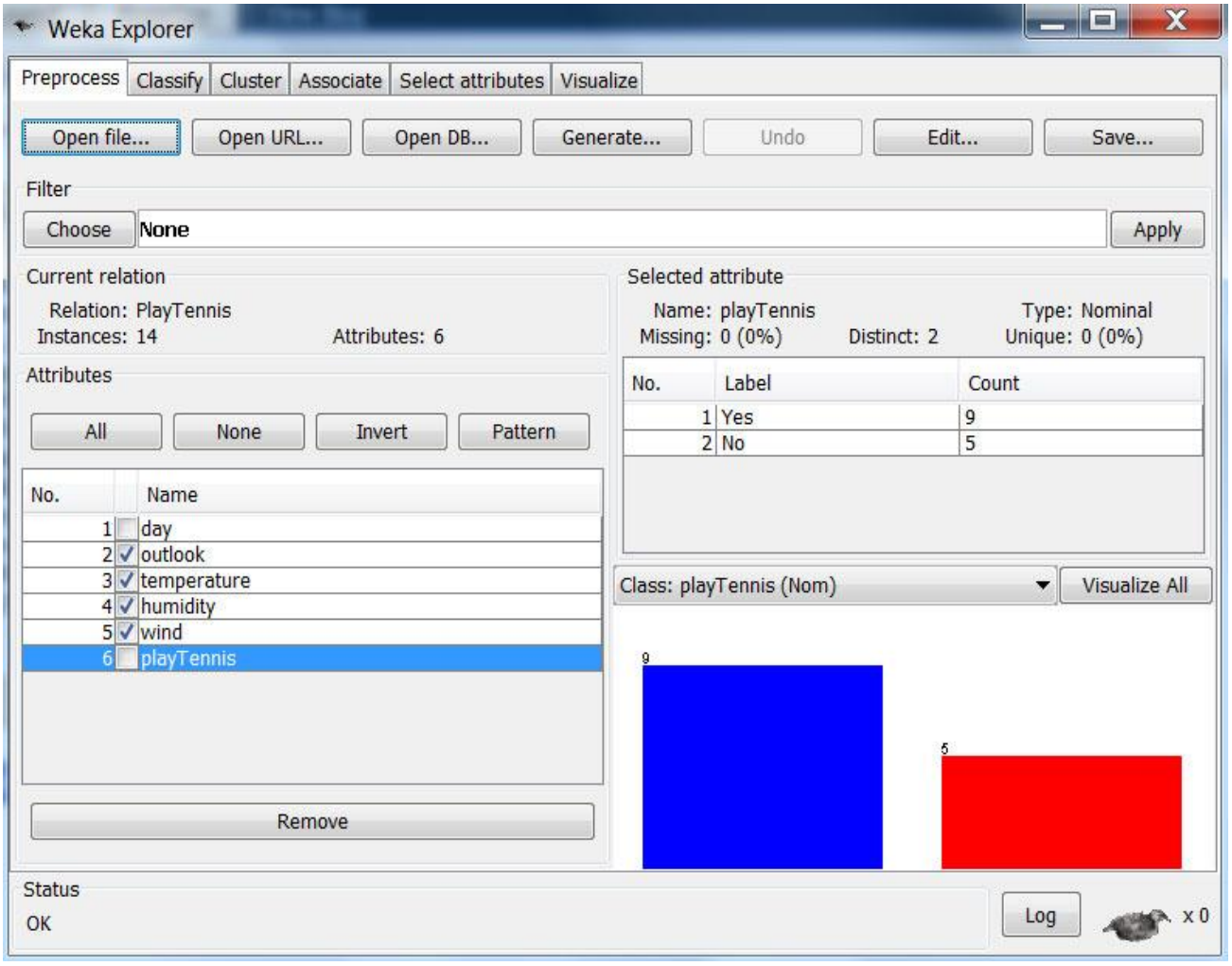
Επιπλέον, έχει αναπτυχθεί μια γεννήτρια δεδομένων/ προσομοιωτής που βασίζεται σε ένα υπάρχον μοντέλο Bayesian δικτύου για δοκιμές και πειραματισμό. Χρησιμοποιώντας τον προσομοιωτή αυτόν, μπορεί κάποιος να επικυρώσει πειραματικά την αιτιώδη υπόθεση που παράγεται από τους αλγορίθμους που βασίζονται στα δίκτυα Bayes.

3.2 Weka

Το WEKA [15] είναι ένα περιβάλλον ανάπτυξης αλγορίθμων και εφαρμογών μηχανικής μάθησης που έχει αναπτυχθεί σε Java και διατίθεται ως ελεύθερο λογισμικό κάτω από την άδεια GNU GPL. Το WEKA έχει αναπτυχθεί και συνεχίζει να αναπτύσσεται στο Πανεπιστήμιο του Waikato στη Νέα Ζηλανδία και το όνομά του έχει δύο σημασίες. Η πρώτη προέρχεται από τα αρχικά των

λέξεων «Waikato Environment for Knowledge Analysis» ενώ η δεύτερη είναι το όνομα ενός πτηνού που ζει μόνο στη Νέα Ζηλανδία

Το WEKA είναι σε θέση να διαβάσει διαφορετικά είδη αρχείων εισόδου, αλλά το πιο συνηθισμένο είναι το ARFF (Attribute- Relation File Format). Πρόκειται για αρχεία κειμένου με συγκεκριμένη μορφή και επέκταση .arff στο όνομά τους.

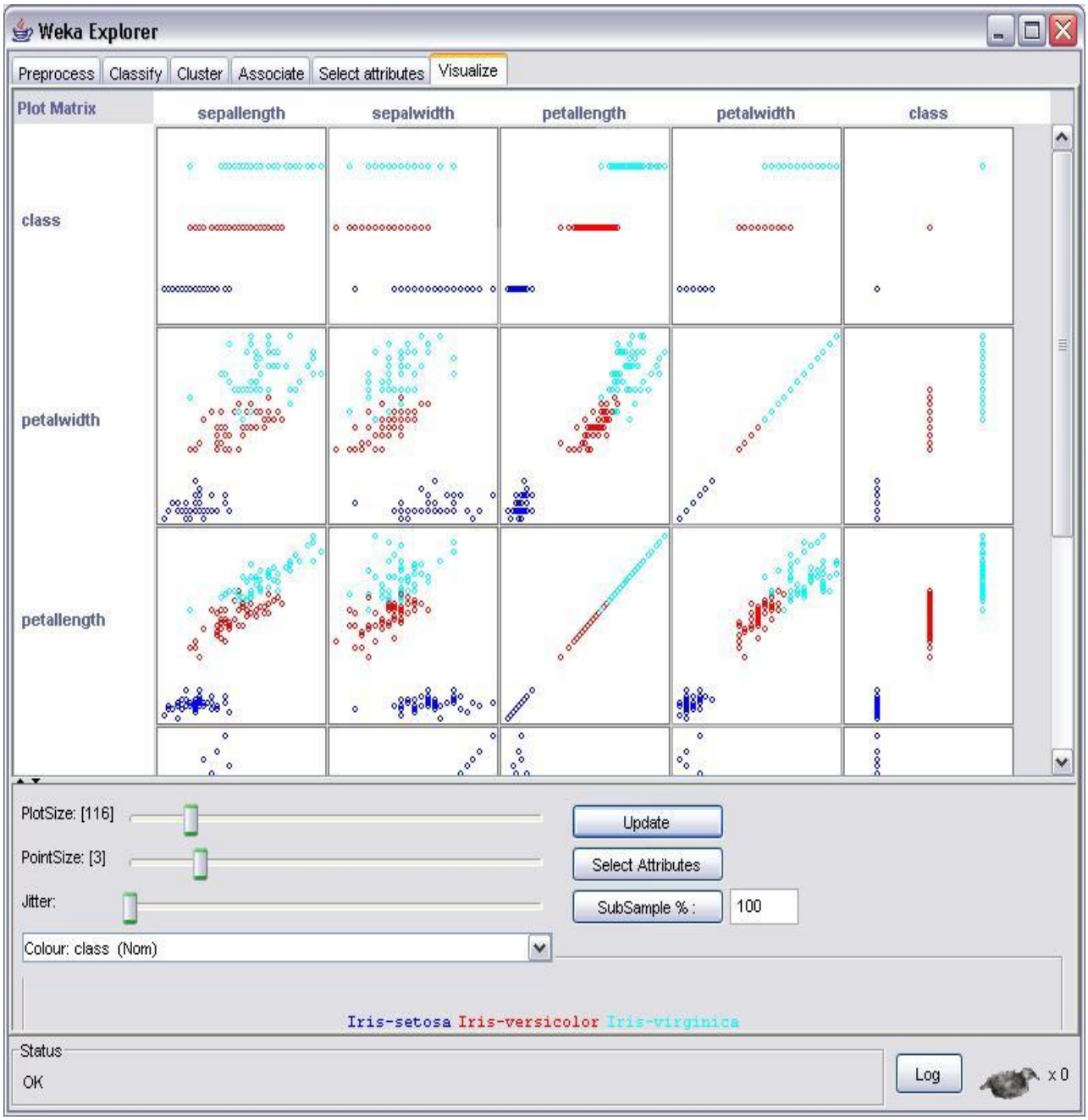


Εικόνα 12 Weka Explorer, Διαδικασίες

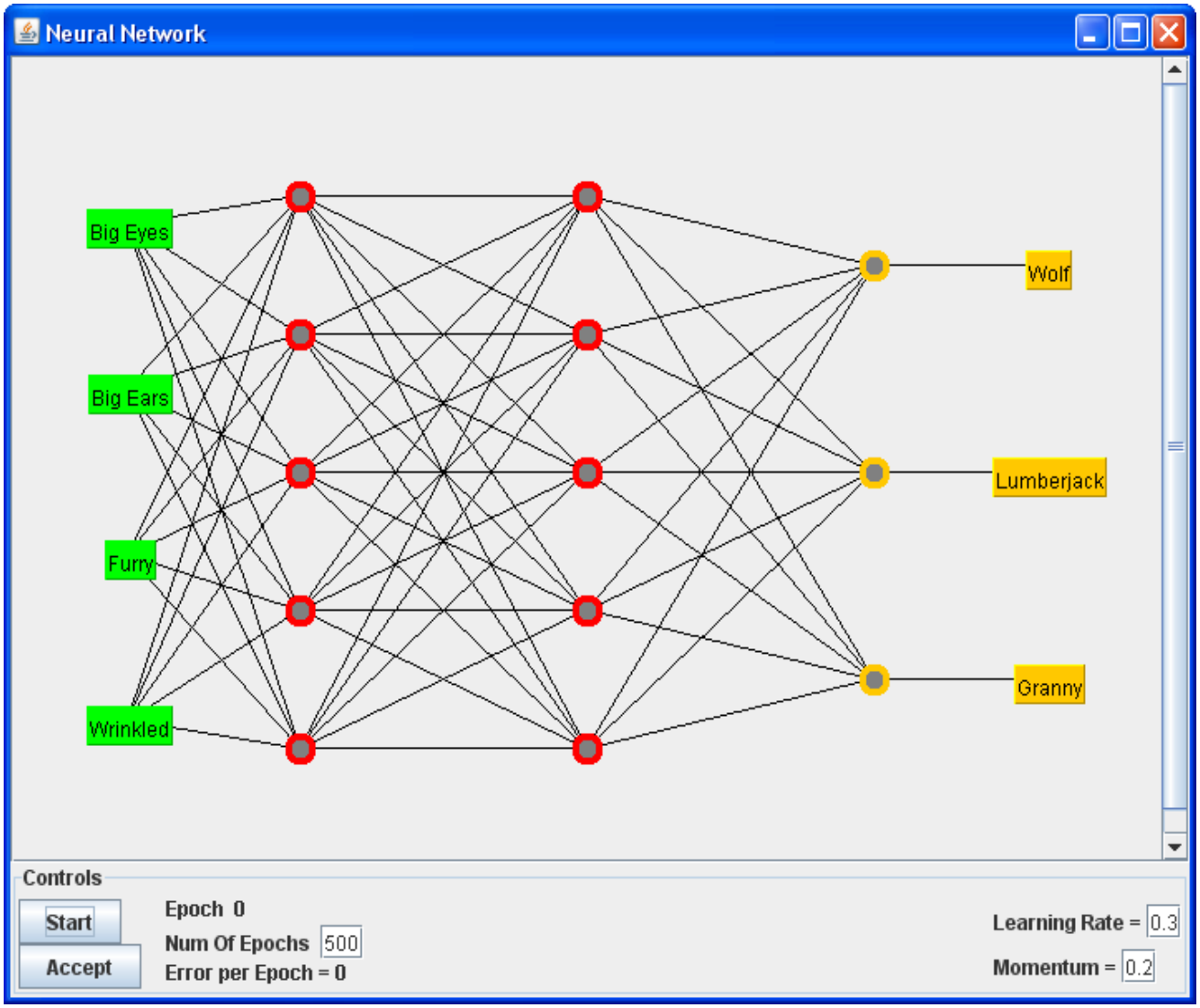
Το Weka [16] αποτελεί μια συλλογή από αλγόριθμους μηχανικής μάθησης προσανατολισμένους για εργασίες σχετικές με την εξόρυξη δεδομένων (data mining). Οι αλγόριθμοι μπορούν είτε να εφαρμοστούν άμεσα σε ένα σύνολο δεδομένων ή να κληθούν από οποιοδήποτε ανεξάρτητο πρόγραμμα γραμμένο σε Java. Μεταξύ άλλων το Weka παρέχει εργαλεία για:

- την απαραίτητη προ- επεξεργασία των δεδομένων
- ταξινόμηση (classification)
- παλινδρόμηση (regression)
- συσταδοποίηση (clustering)

- κανόνες συσχέτισης (association rules)
- οπτικοποίηση (visualization)

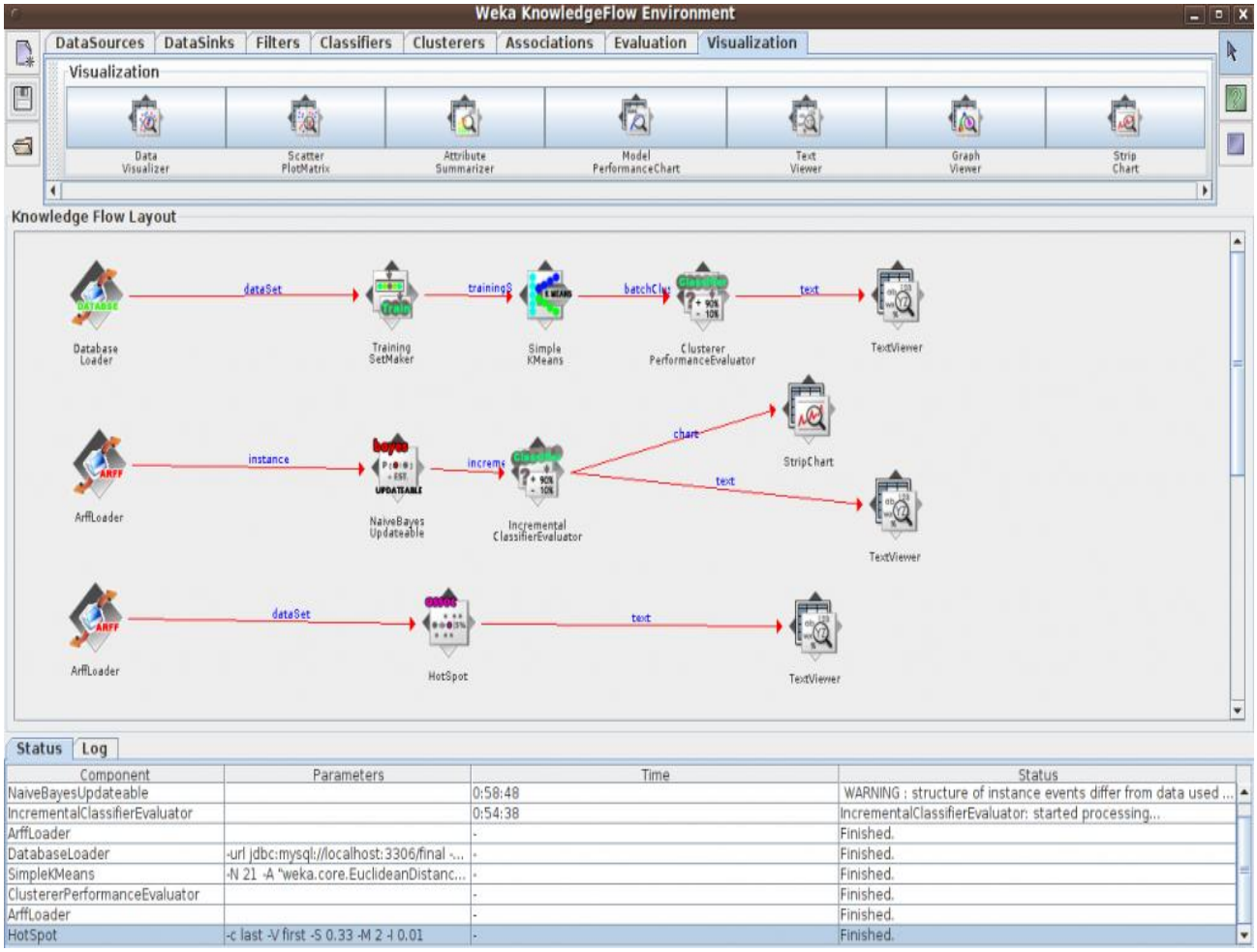


Εικόνα 13 Weka Explorer, οπτικοποίηση



Εικόνα 14 Νευρωνικό Δίκτυο (Neural Network)

Η κεντρική διεπαφή του Weka είναι ο «*Explorer*» και από κει είναι προσβάσιμες σχεδόν όλες οι λειτουργίες του. Φυσικά πρόσβαση στις λειτουργίες του λογισμικού υπάρχει και μέσω γραμμής εντολών αλλά και του «*Knowledge Flow*».



Εικόνα 15 Knowledge Flow περιβάλλον

Με το «*Knowledge Flow*» μπορεί κανείς να οργανώσει μια λειτουργία μέσω γραφικών αντικειμένων του στυλ «drag' n' drop» φτιάχνοντας μια «ροή» λειτουργιών που θα εκτελεστούν από το Weka. Τέλος υπάρχει και η διεπαφή «*Experimenter*» μέσω της οποίας μπορείτε να κάνετε συγκρίσεις στης απόδοση των αλγορίθμων του Weka εφαρμοζόμενων σε μια συλλογή datasets.

3.3 Υλοποίηση κώδικα σε Java

Στα πλαίσια της παρούσας εργασίας υλοποιήθηκε ένα κομμάτι του αλγορίθμου των Koivisto και Sood όπως αυτός προτάθηκε στην εργασία τους και στο άρθρο του Koivisto[17]. Ο αλγόριθμος αυτός υλοποιήθηκε στη γλώσσα προγραμματισμού Java σε περιβάλλον Netbeans 8 με λειτουργικό σύστημα Windows. Το πρόγραμμα προσαρτήθηκε στη βιβλιοθήκη του Weka με τους αλγορίθμους εύρεσης δομής σε Bayesian Networks και κατόπιν ενσωματώθηκε στο εργαλείο AITION που αναλύθηκε παραπάνω.

Οι συναρτήσεις υλοποιήθηκαν με σκοπό την προσαρμογή τους στο πλαίσιο του WEKA, επομένως ακολουθήθηκαν οι οδηγίες προσθήκης ενός νέου εργαλείου μάθησης δομής που παρατίθενται στο εγχειρίδιο χρήσης του WEKA[18]. Σύμφωνα με αυτό, αρχικά δημιουργήθηκε μία κλάση που να προέρχεται από την κλάση *weka.classifiers.bayes.net.search.SearchAlgorithm* και σε αυτήν υλοποιήθηκε η μέθοδος *public void search* στην οποία βρίσκεται ο αλγόριθμος εύρεσης της δομής του δικτύου.

Ο αλγόριθμος του Koivisto είναι ένας ακριβής αλγόριθμος που δίνει τη δυνατότητα για την εύρεση δομής σε ένα Bayesian δίκτυο με πιστότητα μεγαλύτερη από παλαιότερους, προσεγγιστικούς αλγορίθμους. Αυτή η βελτίωση σε αποτέλεσμα και απόδοση επιτυγχάνεται με τη χρήση μιας τεχνικής ανάλογης με την «Μπρος-πίσω» μέθοδο στα κρυφά μοντέλα Markov και με τη χρήση ενός γρήγορο περικομμένου μετασχηματισμού Mobius για την πραγματοποίηση των υπολογισμών.

Σύμφωνα με τον αλγόριθμο αυτό, δεχόμαστε ως είσοδο τα εμπειρικά δεδομένα που έχουμε, δηλαδή τα instances, στα οποία εμφανίζονται δείγματα από τις διάφορες καταστάσεις των κόμβων που εξετάζονται. Η δομή του δικτύου που ψάχνουμε προβάλλει τους υπό-όρους ισχυρισμούς ανεξαρτησίας (conditional independence assertions) ανάμεσα στις μεταβλητές, μέσω ενός κατευθυνόμενου άκυκλου γραφήματος και για να την ορίσουμε θα βρούμε την εκ των υστέρων πιθανότητα κάθε ακμής του. Με βάση την πιθανότητα αυτή, εφόσον κριθεί επαρκώς μεγάλη, σύμφωνα με μια τιμή που έχουμε ορίσει (με βάση τα πειραματικά μας δεδομένα), θα προσθέσουμε την ακμή αυτή στο γράφο. Αυτό πρακτικά σημαίνει ότι υπάρχει εξάρτηση από τον κόμβο-γονέα στον κόμβο-παιδί και επηρεάζοντας τη μία μεταβλητή υπάρχουν επιπτώσεις και για την άλλη. Ορίζουμε το γράφο αυτό ως ένα διάνυσμα $G=(G_1,....,G_n)$ όπου G_i είναι ένα υποσύνολο του συνόλου ευρετηρίου $V=\{1,...,n\}$ και καθορίζει τους «κόμβους πατέρες» του i στο γράφημα.

Τα βήματα του αλγορίθμου είναι τα εξής:

1. Για όλους τους κόμβους $i\in V$ και τα υποσύνολα $G_i\subseteq V-\{i\}$ με $|G_i|\le k$, όπου k είναι ο μέγιστος αριθμός ακμών που μπορεί να έχει ο κόμβος (maximum indegree) υπολογίζουμε το: $\beta_i(G_i):=\rho_i(G_i)\,p(x_i\mid x_{G_i},G_i)\,f_i(G_i)$, όπου $f_i(G_i)$είναι 1 για κάθε i εφόσον αυτό δεν είναι ο κόμβος προορισμού της ακμής.
2. Για όλα τα $i\in V$ και $U_i\subseteq V-\{i\}$ υπολογίζουμε το $a_i(U_i):=q_i(U_i)\sum_{G_i\subseteq U_i}\beta_i(G_i)$
3. Για όλα τα $S\subseteq V$ υπολογίζουμε το $L(S):=\sum_{i\in S}a_i(S-\{i\})L(S-\{i\})$ με $L(\varnothing):=1$

4. Για όλα τα $T \subseteq V$ υπολογίζουμε το $R(T) = \sum_{i \in T} a_i \left(V - T - \{i\}\right) R\left(T - \{i\}\right)$ με $R(\varnothing) = 1$
5. Για όλα τα $v \subseteq V$ (θεωρώντας ως ακμή το ζευγάρι e=(u,v)):
   a. Για όλα τα $G_v \subseteq V - \{i\}$ με $|G_u| \le k$ υπολογίζουμε το $\gamma_v(G_v) := \sum_{G_v \subseteq S \subseteq V - \{i\}} q_v(S) L(S) R\left(V - \{v\} - S\right)$
   b. Για όλα τα $v \subseteq V - \{v\}$ υπολογίζουμε την πιθανότητα του στοιχείου x και της ακμής e σύμφωνα με το: $p(x,e) = \sum_{v \in G_v \subseteq V - \{v\} : |G_v| \le k} \beta_v(G_v) \gamma_v(G_v)$ και δίνουμε ως αποτέλεσμα την εκ των υστέρων πιθανότητα $p(e \mid x) = p(x,e) / L(V)$.

Η μέθοδος «μπρος-πίσω» που αναφέρθηκε αντικατοπτρίζεται στους όρους L(S) και R(S) που αποτελούν τον «εμπρός» και τον «πίσω» υπολογισμό αντίστοιχα. Μπορούμε να χαρακτηρίσουμε το χρόνο εκτέλεσης του παραπάνω αλγορίθμου, με τη συνθήκη ότι ο μέγιστος αριθμός ακμών κάθε κόμβου (k) είναι σταθερός, ως εξής: Το 1ο βήμα χρειάζεται $O\left(n^{k+1}\right)$, το 2ο βήμα $O\left(2^n\right)$, το 3ο και 4ο βήμα $O\left(2^n\right)$ χρόνο. Για κάθε κόμβο $v \subseteq V$ το 5(a), αν χρησιμοποιηθούν οι γρήγοροι περικομμένοι μετασχηματισμοί Mobius, μπορεί να υπολογιστεί σε $O\left(2^n\right)$ και το 5(b) χρειάζεται $O\left(n^k\right)$. Επομένως ολόκληρος ο υπολογισμός χρειάζεται χρόνο $O\left(n2^n\right)$

Στη συνέχεια, στη μέθοδο search που υλοποιήθηκε, ελέγχουμε αν η πιθανότητα που προέκυψε είναι μεγαλύτερη της σταθεράς που έχουμε ορίσει (0.6) και αν αυτό ισχύει και η ακμή δεν περιλαμβάνεται ήδη στο γράφο μας (επειδή μπορεί να έχει περιληφθεί ήδη η ανάποδη ee=(v,u)) την προσθέτουμε σε αυτόν. Τα δεδομένα που προκύπτουν επιστρέφουν ως μια δομή BayesNet ορισμένη από τη βιβλιοθήκη του WEKA και την οποία επεξεργάζεται στη συνέχεια το σύστημα AITION για να προκύψει η επιθυμητή οπτικοποίηση των αποτελεσμάτων.

Για να προστεθεί ο αλγόριθμος αυτός στη βιβλιοθήκη του WEKA πήραμε το .java αρχείο της κλάσης μας καθώς και τα απαραίτητα αρχεία των υποκλάσεων που χρησιμοποιήθηκαν στους υπολογισμούς και τα προσθέσαμε στο weka.jar, αρχείο που δίνεται από το WEKA με σκοπό τη χρησιμοποίηση του java κώδικα που έχει αναπτυχθεί γι αυτό. Η προσθήκη αυτή έγινε στο path: weka/classifiers/bayes/net/search/local στο οποίο βρίσκονται όλοι οι αντίστοιχοι αλγόριθμοι εύρεσης δομής για Bayesian networks.

Τέλος, για να ενσωματωθεί η επιλογή χρήσης του αλγορίθμου από το πρόγραμμα AITION δημιουργήθηκε ένα καινούριο προφίλ αλγορίθμου (αρχείο .profile) με βάση την ήδη υπάρχουσα υλοποίηση παρόμοιων προφίλ για άλλους αλγορίθμους και προστέθηκε στο σύστημα στον κατάλογο: aition/profiles/learning. Στο προφίλ αυτό περιλαμβάνεται η περιγραφή του

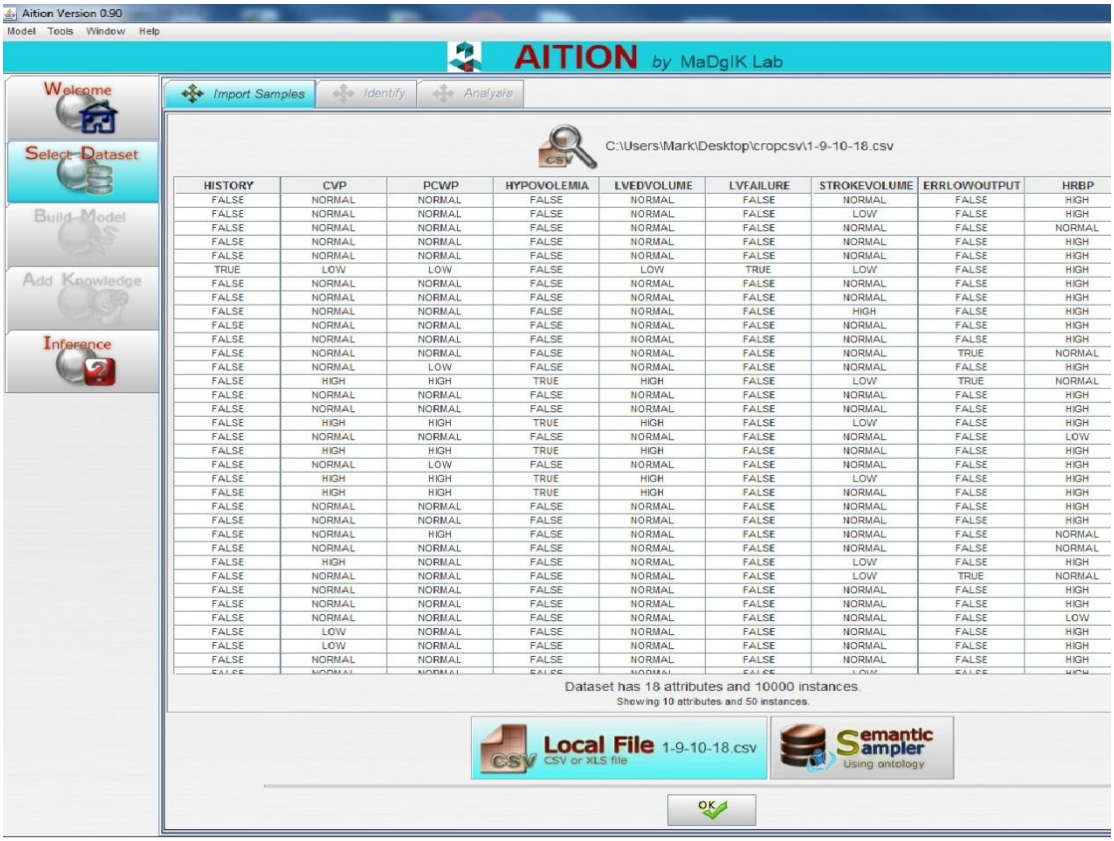
αλγορίθμου, η τοποθεσία του μέσα στη βιβλιοθήκη του WEKA καθώς και προεπιλογές για τη σωστή εκτέλεση του αλγορίθμου. Στον ίδιο κατάλογο προστέθηκε ένα επιπλέον προφίλ για τον (ήδη υλοποιημένο στο σύστημα WEKA) αλγόριθμο K2 με σκοπό την περίληψή του στο σύστημα και τη χρήση του για σκοπούς δοκιμών.

4. ΠΕΙΡΑΜΑΤΑ ΚΑΙ ΑΞΙΟΛΟΓΗΣΗ

4.1 Πειραματικά δεδομένα

Προκειμένου να αξιολογηθεί σωστά ο αλγόριθμος εύρεσης δομής που υλοποιήθηκε, πραγματοποιήθηκαν πειραματικές εκτελέσεις του και παρατήρηση-ανάλυση των αποτελεσμάτων που προέκυψαν.

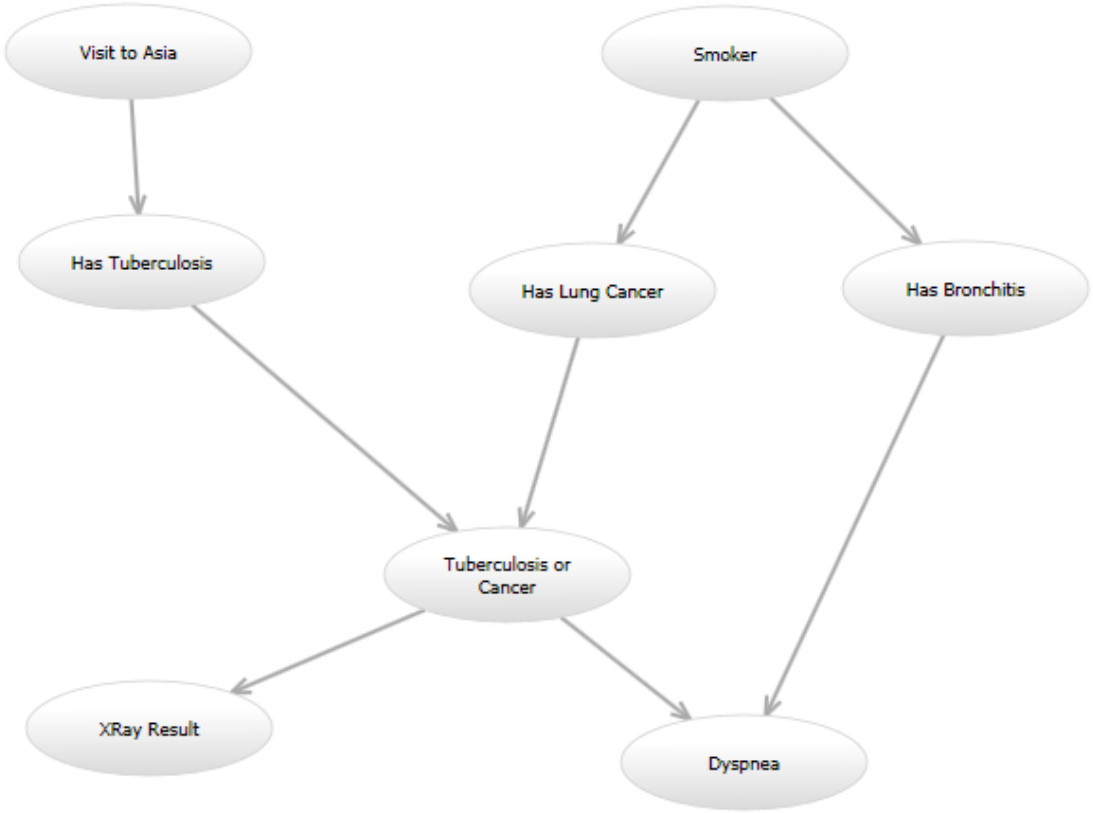
Τα δεδομένα τα εισάγαμε στο σύστημα AITION. Αποτελούνταν από αρχεία δεδομένων .csv όπου στην πρώτη σειρά βρίσκονταν οι μεταβλητές διαχωρισμένες με κόμματα και στις επόμενες σειρές ήταν οι καταστάσεις των μεταβλητών από τυχαία παραδείγματα με βάση το δίκτυο στο οποίο ανήκουν.



Εικόνα 16 Εισαγωγή δεδομένων στο σύστημα AITION

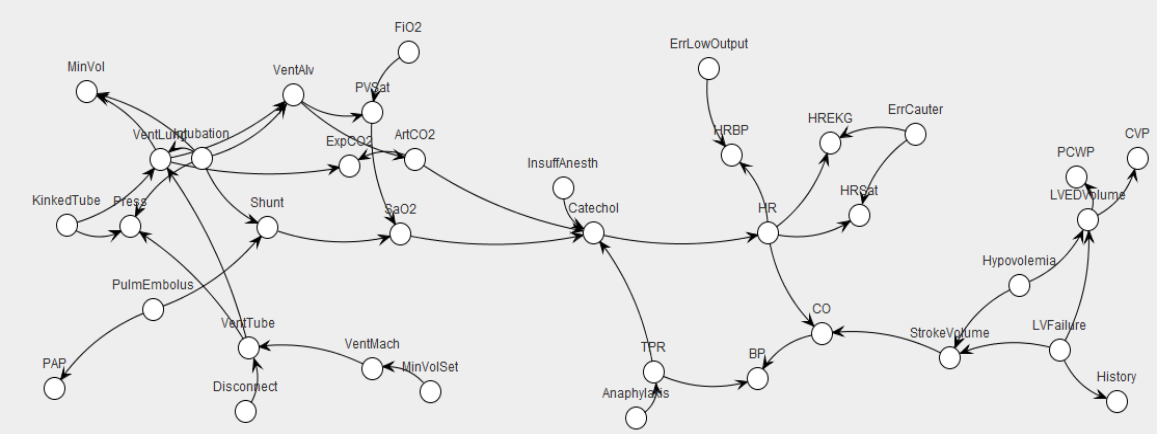
Οι εκτελέσεις αυτές βασίστηκαν κυρίως σε 2 ομάδες δεδομένων. Η πρώτη ομάδα βασίζεται στο Bayesian δίκτυο asia (όπως φαίνεται στην εικόνα 17). Ο γράφος του δικτύου αυτού περιλαμβάνει 8 κόμβους που αποτελούν τις μεταβλητές και 8 ακμές που δηλώνουν τις συσχετίσεις μεταξύ των μεταβλητών αυτών. Οι καταστάσεις των μεταβλητών είναι δυαδικές, δηλαδή κάθε μεταβλητή μπορεί να πάρει είτε την τιμή True, είτε την τιμή False. Ο

αριθμός των μέγιστων ακμών που μπορεί να έχει ένας κόμβος στο δίκτυο αυτό είναι 4.



Εικόνα 17 Το δίκτυο ASIA

Η δεύτερη ομάδα δεδομένων που χρησιμοποιήθηκε βασίζεται στο Bayesian δίκτυο alarm (όπως φαίνεται στην εικόνα 18). Σε αυτό το δίκτυο έχουμε 37 κόμβους-μεταβλητές και 46 συσχετίσεις μεταξύ τους. Αποτελεί λοιπόν ένα αρκετά πιο περίπλοκο δίκτυο και λόγο του μεγάλου αριθμού μεταβλητών έπρεπε να χωριστεί σε επιμέρους κομμάτια για την ορθότερη ανάλυσή του. Ο διαχωρισμός αυτός αναλύεται στη συνέχεια. Οι μεταβλητές του δικτύου αυτού είχαν διαφορετικό αριθμό καταστάσεων η κάθε μία όπως για παράδειγμα: low-normal-high ή true-false. Ο μέγιστος αριθμός ακμών σε κάποιον κόμβο είναι 5 σε αυτήν την περίπτωση.



Εικόνα 18 Το δίκτυο ALARM

4.2 Τμηματοποίηση δεδομένων

Τα δεδομένα που αντιστοιχούν στο δίκτυο asia λόγω του ότι ο αριθμός των μεταβλητών του δικτύου είναι μικρός (8) αξιοποιήθηκαν ολοκληρωμένα χωρίς να απαιτούνται περαιτέρω τροποποιήσεις τους, με βάση των αριθμό των κόμβων. Τα συνολικά δεδομένα αποτελούνταν από 10000 εγγραφές και πραγματοποιήθηκαν ξεχωριστές μετρήσεις για 200, 500, 1000, 2500, 5000, 8000 και 10000 εγγραφές. Τα δείγματα αυτά λήφθηκαν τυχαία μέσα από το αρχείο των 10000 εγγραφών με σκοπό να μην επηρεαστούν τα αποτελέσματα από τυχόν ιδιομορφίες της κατανομής του συνολικού δείγματος.

Το δίκτυο Alarm, ωστόσο, ήταν αρκετά μεγαλύτερο και λόγο περιορισμών δεν ήταν δυνατή η χρήση των δεδομένων χωρίς να χωριστούν σε μικρότερα κομμάτια. Ο αριθμός των κόμβων (37) ήταν μεγαλύτερος από το μέγιστο όριο (19) που έχει τη δυνατότητα ο αλγόριθμός μας να επεξεργαστεί με επιτυχία στο σύστημα που δοκιμάστηκε. Για το λόγο αυτό τα δεδομένα μας χωρίστηκαν σε υποομάδες των 9 κόμβων. Δημιουργήθηκαν έτσι 4 ομάδες και απέμενε μία μεταβλητή η οποία προστέθηκε στην τελευταία υποομάδα. Στη συνέχεια κάθε υποομάδα δοκιμάστηκε σε κάθε δυνατό συνδυασμό με τις υπόλοιπες. Αναλυτικότερα οι κόμβοι 1-9 δοκιμάστηκαν με τους 10-18, ύστερα με τους 19-27 και κατόπιν με τους 28-37. Αντίστοιχα στη συνέχεια, οι κόμβοι 10-18 με τους 19-27 και με τους 28-37 και τέλος οι 19-27 με τους 28-37. Με τον τρόπο αυτό ελέγχθηκε αν εμφανίζονται σωστά οι συσχετίσεις μεταξύ οποιουδήποτε κόμβου χωρίς να αποκλείεται κάποιος συνδυασμός. Στη συνέχεια έγινε συλλογή των αποτελεσμάτων της κάθε ξεχωριστής μέτρησης και παρουσιάζονται τα συνολικά αποτελέσματα για το δίκτυο.

Όπως και στο δίκτυο asia που προαναφέρθηκε, έτσι και στο alarm υπήρχε ένα σύνολο από 10000 εγγραφές και πραγματοποιήθηκαν ξεχωριστές μετρήσεις για 1000, 5000 και 10000 εγγραφές. Αντίστοιχα, και σε αυτές τις εκτελέσεις, τα επιμέρους δείγματα λήφθηκαν τυχαία.

4.3 Αποτελέσματα με βάση το γράφο

Η πρώτη σειρά μετρήσεων που πραγματοποιήθηκε αφορούσε την ορθότητα του γράφου. Για την καλύτερη κατανόηση των αποτελεσμάτων, πέρα από του αλγορίθμου που εξετάζεται (Rebel) πραγματοποιήθηκαν μετρήσεις και με άλλους 4 αλγορίθμους οι οποίοι βρίσκονται στις βιβλιοθήκες του WEKA και έχουν επίσης ενσωματωθεί στο σύστημα AITION. Παρακάτω θα δοθεί μια σύντομη περιγραφή του κάθε έναν από αυτούς.

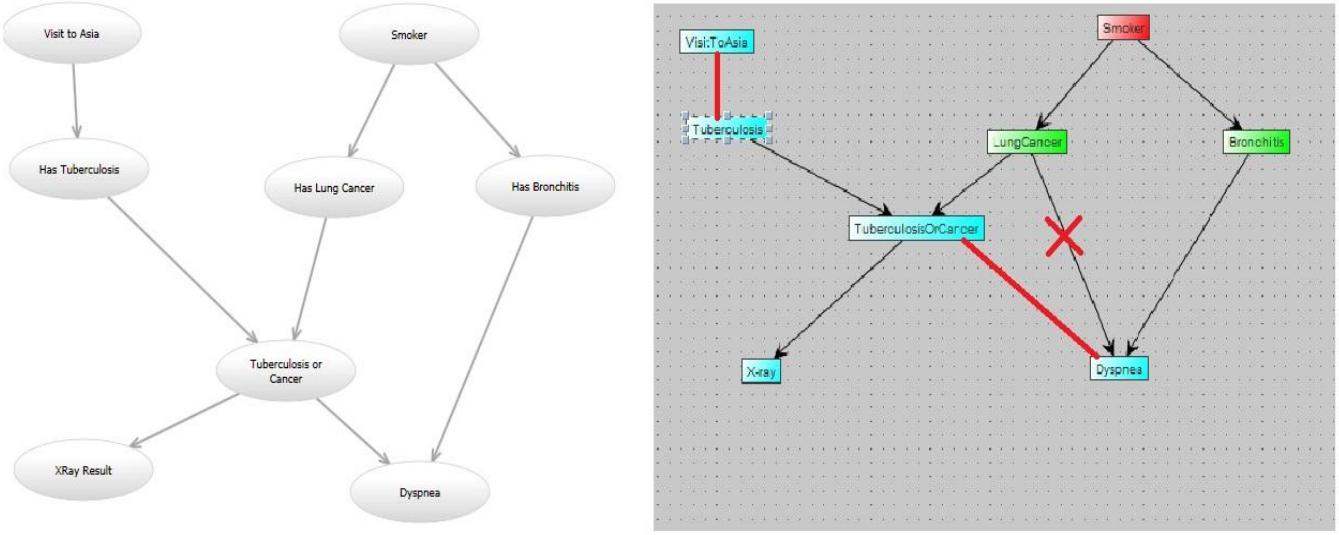
Ο πρώτος είναι ο hill-climbing αλγόριθμος. Ο αλγόριθμος αυτός προσθέτει, αφαιρεί και αντιστρέφει ακμές στο γράφο με βάση μια μέθοδο βαθμολόγησης και στο τέλος προκύπτει ο γράφος με την καλύτερη συνολική βαθμολογία. Παρουσιάζει καλύτερα αποτελέσματα από άλλους σε προβλήματα που έχουν γράφους παρόμοιους με το ASIA δίκτυο.

Ο δεύτερος είναι ο αλγόριθμος simulated-annealing[1]. Αυτός χρησιμοποιεί μία μέθοδο εύρεσης δομής προσομοιωμένης ανόπτησης που είναι γενικού σκοπού με σκοπό να βρει ένα γράφο με καλή βαθμολόγηση.

Ο τρίτος είναι ο αλγόριθμος TAN[2]. Ο αλγόριθμος αυτός βρίσκει το δέντρο με το μεγαλύτερο εκτεινόμενο βάρος και επιστρέφει ένα Naive Bayesian δίκτυο επαυξημένο με το δέντρο αυτό.

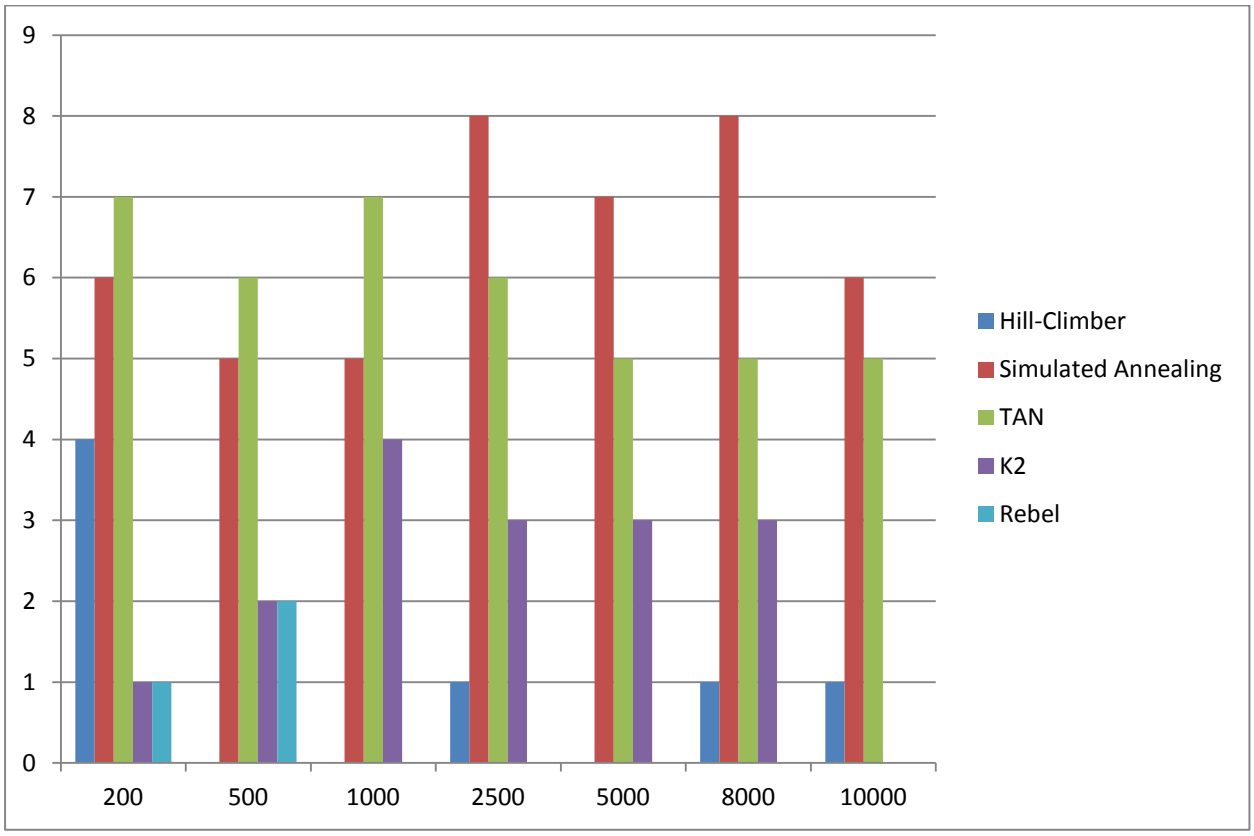
Ο τέταρτος αλγόριθμος ονομάζεται K2[3]. Είναι ένας αλγόριθμος που λειτουργεί με τον ίδιο τρόπο όπως και ο hill-climbing με τη διαφορά ότι ο K2 δεν επηρεάζεται από τη σειρά με την οποία είναι διατεταγμένες οι μεταβλητές.

Για κάθε έναν από αυτούς τους αλγόριθμους εκτελέστηκε αρχικά για το ASIA δίκτυο και ύστερα για το ALARM το πρόγραμμα AITION και έγινε σύγκριση του γράφου που προέκυψε ως αποτέλεσμα με τον ορθό γράφο του αντίστοιχου Bayesian δικτύου. Παρατηρήθηκαν έτσι οι πλεονάζουσες αλλά και οι ελλιπείς συσχετίσεις που βρέθηκαν από κάθε ξεχωριστό αλγόριθμο.

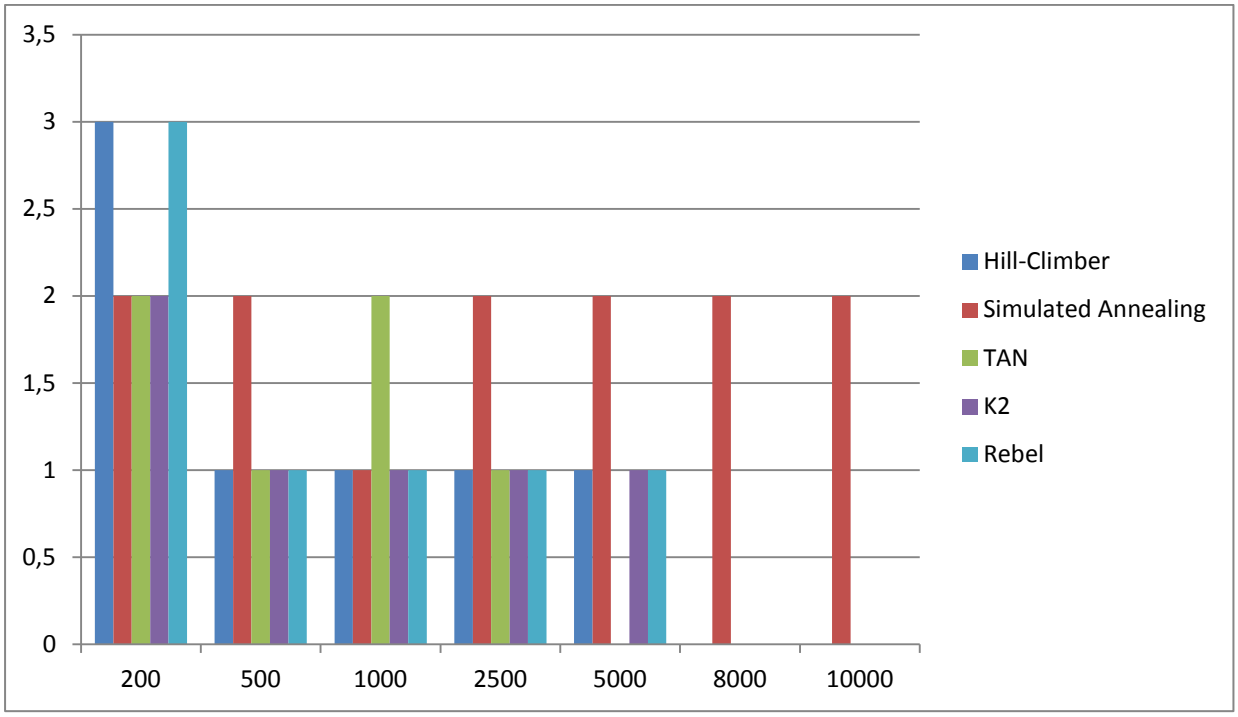


Εικόνα 19 Σύγκριση γράφου που προέκυψε από τον αλγόριθμο K2 για 200 εγγραφές με τον ορθό γράφο. Με κόκκινες γραμμές επισημαίνονται οι ακμές που απουσιάζουν και με κόκκινο Χ έχουν τονιστεί οι ακμές που δεν υπάρχουν στον ορθό γράφο.

Για το δίκτυο ASIA οι μετρήσεις αυτές έγιναν για 200, 500, 1000, 2500, 5000, 8000 και 10000 εγγραφές και τα αποτελέσματα συνοψίζονται στις εικόνες 20 και 21 όπου παρουσιάζονται οι επιπλέον και οι λιγότερες ακμές.



Εικόνα 20 Ο αριθμός των παραπάνω ακμών που βρήκε ο κάθε αλγόριθμος σε σχέση με τον ορθό γράφο του ASIA δικτύου για 200, 500, 1000, 2500, 5000, 8000, 10000 εγγραφές

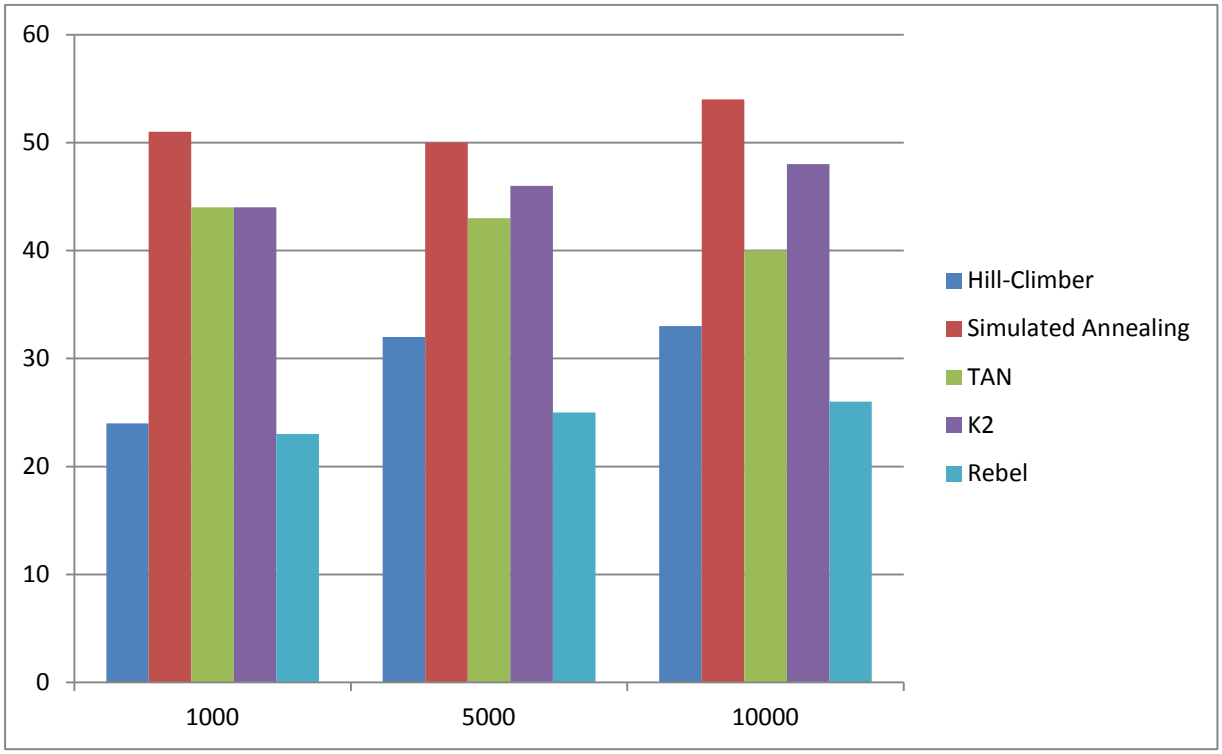


Εικόνα 21 Ο αριθμός των ακμών που δεν βρήκε ο κάθε αλγόριθμος σε σχέση με τον ορθό γράφο του ASIA δικτύου για 200, 500, 1000, 2500, 5000, 8000, 10000 εγγραφές

Σύμφωνα με τις παραπάνω εικόνες γίνεται εμφανής η ανωτερότητα του αλγορίθμου του Koivisto σε σχέση με του ήδη υπάρχοντες αλγορίθμους όσον αφορά την ακρίβεια στο γράφο. Γίνεται εμφανές ότι οι πλεονάζουσες ακμές με τον αλγόριθμο αυτό είναι σε κάθε περίπτωση λιγότερες ή ίσες από τους υπόλοιπους αλγορίθμους. Πιο συγκεκριμένα εμφανίζονται 1 και 2 ακμές και μόνο σε περίπτωση που τα δεδομένα μας είναι λιγότερα από 1000, ενώ οι υπόλοιποι αλγόριθμοι, έχουν αρκετές παραπάνω ακμές. Η μόνη εξαίρεση σε αυτό είναι ο hill-climber αλγόριθμος που συγκεκριμένα για το ASIA πρόβλημα εμφανίζει πολύ καλά αποτελέσματα τα οποία και πάλι δεν ξεπερνούν σε ακρίβεια αυτά του Rebel.

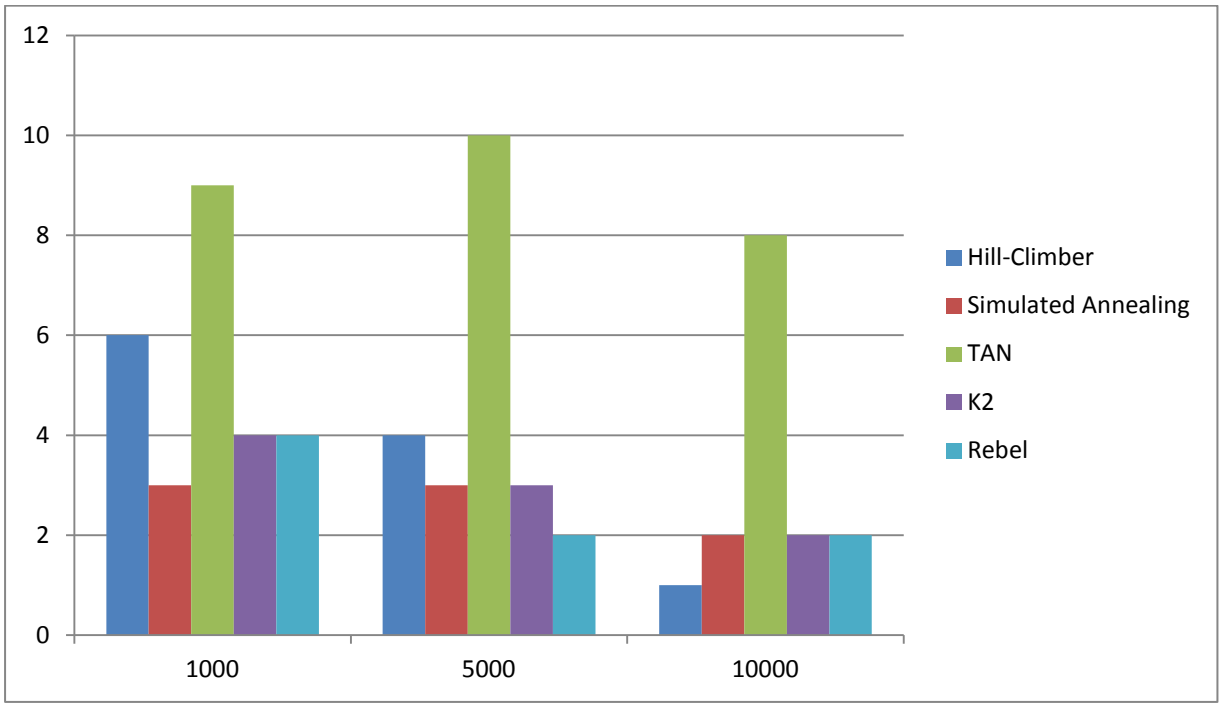
Στην εικόνα 21 παρατηρούνται επίσης πολύ καλά αποτελέσματα του αλγορίθμου σε σχέση με τις συσχετίσεις που δεν κατάφερε να εντοπίσει. Πέρα από την περίπτωση με πολύ λίγα δεδομένα (<500), και πάλι είναι ο βέλτιστος, σαν σύνολο, αλγόριθμος σε σύγκριση με τους υπόλοιπους 4. Συνοψίζοντας λοιπόν, παρατηρείται ότι στο δίκτυο ASIA τα πιο ακριβή αποτελέσματα παρουσιάζονται με τον αλγόριθμο που υλοποιήθηκε και ο γράφος σαν αποτέλεσμα είναι ο πιο κοντινός με τον ορθό. Ιδιαίτερα με μεγάλο όγκο δεδομένων ο γράφος που ανακαλύπτει είναι εντελώς σωστός.

Στις παρακάτω εικόνες (22 και 23) εμφανίζονται τα αντίστοιχα αποτελέσματα με τους ίδιους αλγορίθμους για το δίκτυο ALARM, αφού αθροίστηκαν μετά το διαχωρισμό των δεδομένων σε υποομάδες των 9 όπως αναφέρθηκε και παραπάνω.



Εικόνα 22 Ο αριθμός των παραπάνω ακμών που βρήκε ο κάθε αλγόριθμος σε σχέση με τον ορθό γράφο του ALARM δικτύου για 1000, 5000, 10000 εγγραφές

Στην εικόνα 22 παρατηρείται ότι για όλα τα μεγέθη εγγραφών οι πλεονάζουσες ακμές που βρήκε ο αλγόριθμος που εξετάζεται είναι λιγότερες από οποιονδήποτε άλλο από τους 4 αλγορίθμους. Η μεγαλύτερη βελτίωση υπάρχει στην εύρεση δομής με τις πιο πολλές εγγραφές, παρόλα αυτά, έστω και με μόνο 1000 τα αποτελέσματα που παρουσιάζονται είναι καλύτερα. Ο αλγόριθμος Simulated Annealing εμφανίζει τα χειρότερα αποτελέσματα με πολλές παραπανίσιες συσχετίσεις δεδομένων και καθίσταται ο λιγότερο χρήσιμος, ενώ αντίθετα ο Hill-Climber είναι πολύ κοντά στον αλγόριθμο Rebel και τα αποτελέσματά του είναι αρκετά ακριβή.



Εικόνα 23 Ο αριθμός των ακμών που δεν βρήκε ο κάθε αλγόριθμος σε σχέση με τον ορθό γράφο του ALARM δικτύου για 1000, 5000, 10000 εγγραφές

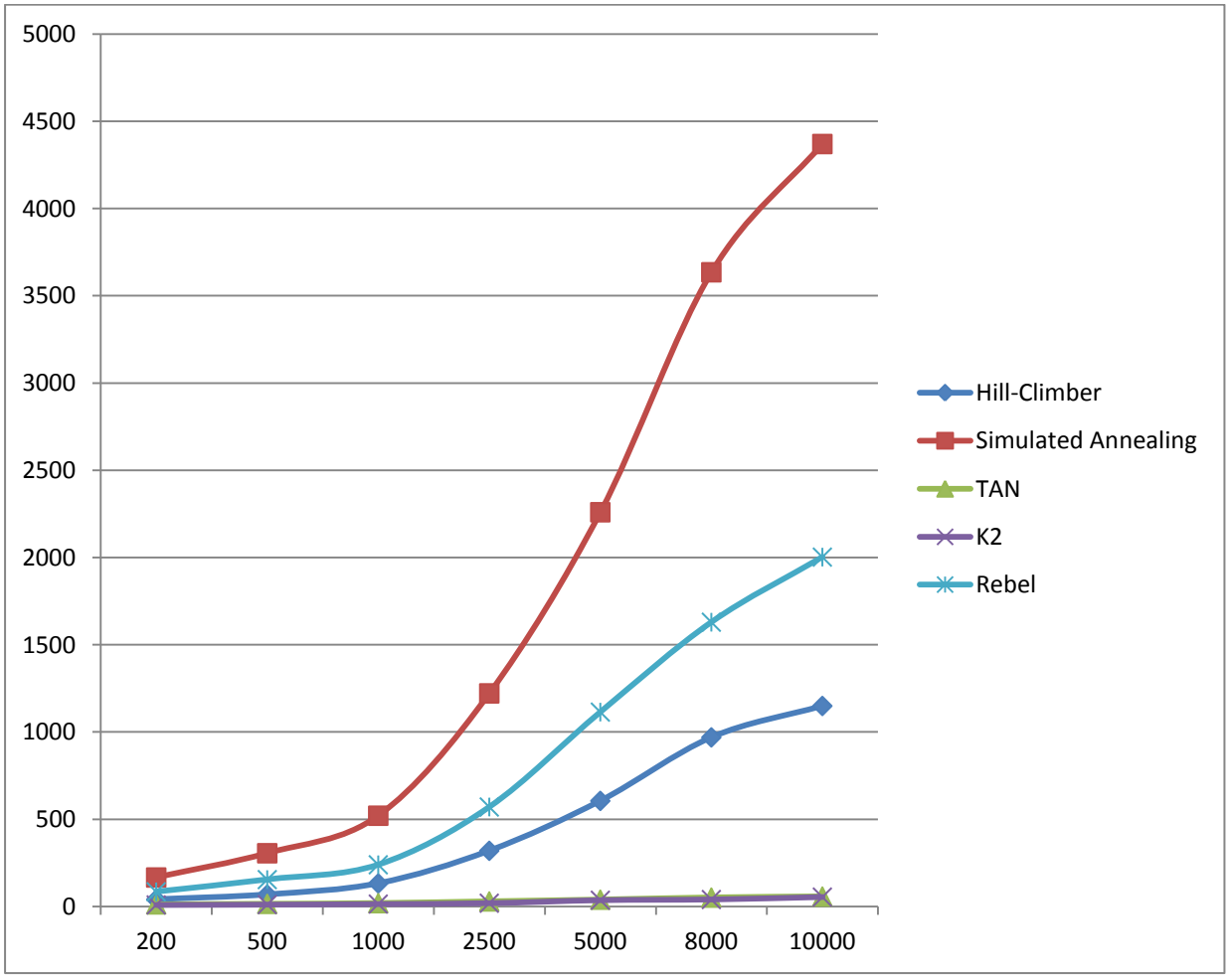
Στην εικόνα 23 παρουσιάζονται οι συσχετίσεις που δεν βρέθηκαν για το γράφο του ALARM δικτύου. Ο αλγόριθμος του Koivisto εμφανίζει και εδώ αρκετά ικανοποιητικά αποτελέσματα. Ο αλγόριθμος Hill-Climber έχει λιγότερες ελλείψεις στις 10000 εγγραφές αλλά παρόλα αυτά με λιγότερα δεδομένα έχει αρκετά περισσότερες, με αποτέλεσμα να μην συμφέρει η χρήση του σε αυτές τις περιπτώσεις. Αντίθετα, ο Simulated Annealing είναι ελαφρώς πιο αποδοτικός όσον αφορά τις ελλείψεις με μικρό όγκο δεδομένων αλλά ο Rebel βελτιώνεται με την αύξηση των δεδομένων. Το γεγονός αυτό σε συνδυασμό με τον πολύ αυξημένο αριθμό περισσευούμενων ακμών που παρατηρήθηκε οδηγεί στο συμπέρασμα ότι η χρήση του Rebel είναι αποδοτικότερη.

Συνοψίζοντας, με βάση τα πειράματα που εκτελέσθηκαν και για τα δύο δίκτυα που εξετάστηκαν γίνεται φανερό αυτό που υποστηρίζεται και από τον Koivisto ότι ο αλγόριθμος για ακριβή εύρεση της δομής ενός Μπεϋζιανού δικτύου είναι σαφώς ανώτερος από τους προσεγγιστικούς αλγορίθμους και ο γράφος πολύ πιο κοντά στον ορθό γράφο του δικτύου.

4.4 Αποτελέσματα με βάση το χρόνο

Η δεύτερη σειρά μετρήσεων αφορούσε το χρόνο εκτέλεσης του κάθε αλγορίθμου και τη σύγκριση μεταξύ τους. Η εκτέλεση των αλγορίθμων πραγματοποιήθηκε για όλες τις μετρήσεις στον ίδιο υπολογιστή για την επιτυχή σύγκριση των αποτελεσμάτων. Το λειτουργικό σύστημα ήταν 32-bit Windows 7 με μνήμη RAM 3GB και επεξεργαστή Intel Core i7 στα 2.67GHz.

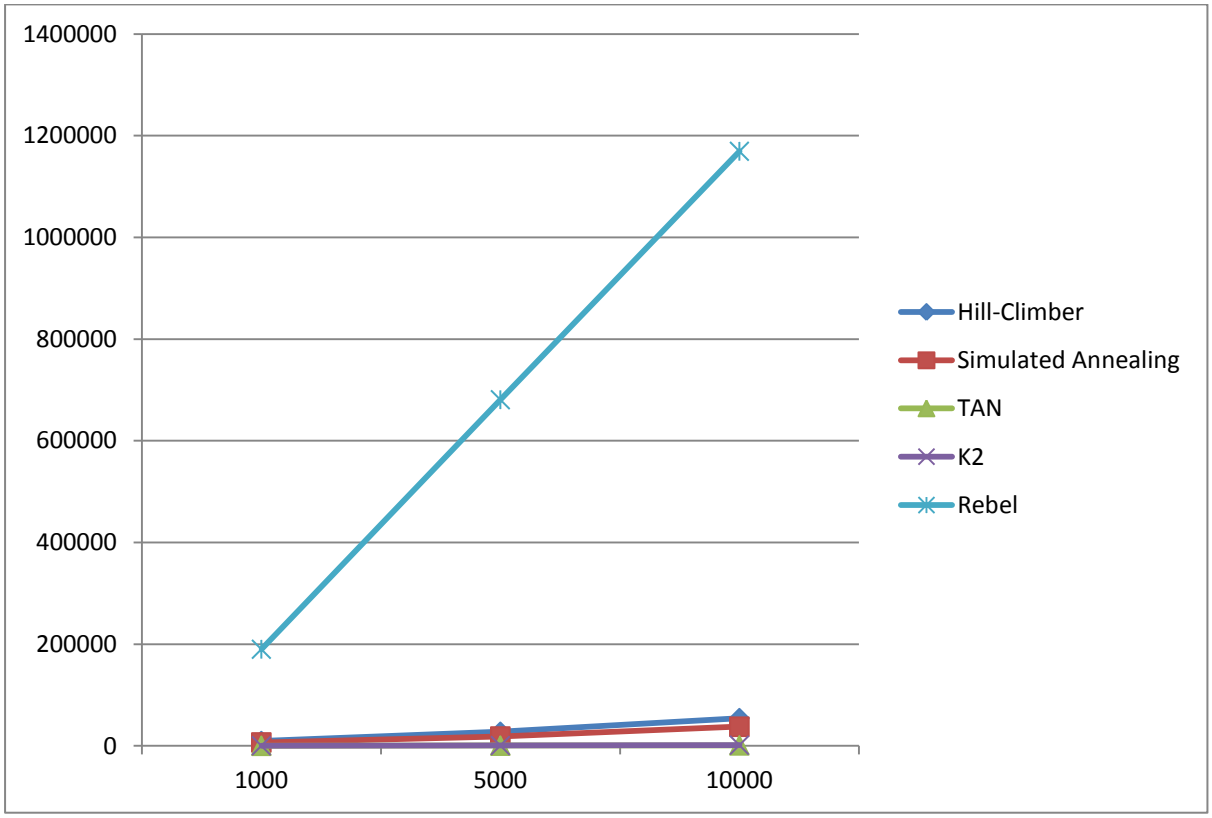
Στο δίκτυο ASIA τα αποτελέσματα που παρουσιάζονται είναι ο χρόνος που χρειάστηκε ο κάθε αλγόριθμος να βρει ολόκληρη τη δομή του δικτύου. Παρόλα αυτά στο δίκτυο ALARM καθότι ο αριθμός των μεταβλητών ήταν μεγαλύτερος από τις δυνατότητες των αλγορίθμων να επεξεργαστούν, τα δεδομένα χωρίστηκαν με τον τρόπο που προαναφέρθηκε και ο χρόνος που παρουσιάζεται είναι το άθροισμα των επιμέρους χρόνων εκτέλεσης για το κάθε κομμάτι και δεν είναι αντιπροσωπευτικό του χρόνου που θα χρειαζόταν ο κάθε αλγόριθμος για να βρει ολόκληρη τη δομή (αν κάτι τέτοιο ήταν εφικτό).



Εικόνα 24 Χρόνοι εκτέλεσης (σε ms) 5 αλγορίθμων για την εύρεση της δομής του δικτύου ASIA με δεδομένα 200, 500, 1000, 2500, 5000, 8000 και 10000 εγγραφών

Στην εικόνα 24 παρουσιάζονται οι χρόνοι εκφρασμένοι σε ms των αλγορίθμων που εξετάστηκαν. Οι χρόνοι αυτοί είναι πραγματικοί χρόνοι εκτέλεσης με

διαφορετικούς αριθμούς εγγραφών. Οι πλέον αποδοτικοί αλγόριθμοι όσον αφορά το χρόνο σε αυτήν την περίπτωση είναι οι αλγόριθμοι K2 και TAN. Όσο τα δεδομένα αυξάνονταν οι χρόνοι τους παρέμεναν σχεδόν γραμμικοί και πολύ χαμηλοί. Αντίθετα εντελώς, οι χρόνοι του simulated annealing για το παράδειγμα αυτό με τις 8 μεταβλητές ήταν ιδιαίτερα μεγάλοι με σχεδόν εκθετική αύξηση, ενώ σε ενδιάμεση κατάσταση βρίσκονται οι χρόνοι του Rebel και του Hill-Climber που και αυτοί αυξάνονται ιδιαίτερα όσο μεγαλώνει το πλήθος των δεδομένων αλλά όχι στο βαθμό του Simulated Annealing.



Εικόνα 25 Χρόνοι εκτέλεσης (σε ms) 5 αλγορίθμων για την εύρεση της δομής του δικτύου ALARM με δεδομένα 1000, 5000 και 10000 εγγραφών

Οι χρόνοι εκτέλεσης (σε ms) των εξεταζόμενων αλγορίθμων για το δίκτυο ALARM παρουσιάζονται στην εικόνα 25. Σε αυτήν, γίνεται φανερό ότι οι 4 προσεγγιστικοί αλγόριθμοι έχουν συντριπτικά χαμηλότερους χρόνους από τον ακριβή αλγόριθμό του Koivisto. Αυτό ισχύει για όλα τα μεγέθη δεδομένων που χρησιμοποιήθηκαν και οφείλεται στον πολύ μεγάλο αριθμό κόμβων του προβλήματος ALARM το οποίο αυξάνει εκθετικά το χρόνο για κάθε παραπάνω κόμβο που εμφανίζεται. Με αυτόν τον τρόπο εξηγείται και η διαφορά στην απόδοση σε σχέση με αντίστοιχα μεγέθη δεδομένων μεταξύ του ASIA και του ALARM δικτύου. Ενώ στο πρώτο οι χρόνοι εκτέλεσης μεταξύ όλων των αλγορίθμων ήταν συγκρίσιμοι, στο παράδειγμα του ALARM δεν συμβαίνει το ίδιο. Η αύξηση του χρόνου σε σχέση με το πλήθος των εγγραφών είναι σαφώς καλύτερη από το πρόβλημα ASIA αλλά τα μεγέθη πλέον των χρόνων έχουν αλλάξει δραματικά.

4.5 Αποτελέσματα

Σύμφωνα με τις παραπάνω πληροφορίες που λάβαμε από τις εκτελέσεις των αλγορίθμων προκύπτουν κάποια πολύ χρήσιμα αποτελέσματα. Αρχικά, παρατηρείται ότι σε δίκτυα με μικρό αριθμό μεταβλητών όπως είναι το πρόβλημα ASIA ο ακριβής αλγόριθμος του Koivisto παρουσιάζει βέλτιστα αποτελέσματα σε σχέση με τους υπόλοιπους αλγορίθμους και η αύξηση στη χρονική διάρκεια εκτέλεσης δεν είναι αρκετή ώστε να αποτρέψει τη χρήση του. Τα αποτελέσματα είναι αρκετά ακριβή ακόμα και με μικρό αριθμό δεδομένων και ο γράφος που προκύπτει είναι πολύ κοντά στον ορθό γράφο του δικτύου. Όσο τα δεδομένα αυξάνονται παράγονται διαρκώς καλύτερα αποτελέσματα. Η χρήση του ενδείκνυται απέναντι στους άλλους 4 αλγορίθμους οι οποίοι μπορεί να εμφανίζουν μικρότερους χρόνους αλλά δεν ικανοποιούν με την πιστότητά τους.

Στην περίπτωση προβλημάτων με πολλές μεταβλητές ωστόσο η καταλληλότητα του αλγορίθμου εξαρτάται και από τις επιθυμίες του χρήστη. Τα αποτελέσματα που προκύπτουν είναι και πάλι σαν σύνολο καλύτερα από των υπόλοιπων, προσεγγιστικών, αλγορίθμων αλλά η διαφορά στο χρόνο εκτέλεσης είναι ιδιαίτερα υψηλή. Όταν οι απαιτήσεις του χρήστη αφορούν την πιστότητα τότε η «θυσία» του χρόνου είναι προσοδοφόρα και η χρήση του αλγορίθμου Rebel είναι επιθυμητή. Παρόλα αυτά η διαφορά στην ποιότητα των αποτελεσμάτων ίσως να μην είναι επαρκής όταν υπάρχει ανάγκη για πιο άμεσα αποτελέσματα και να προτιμηθεί κάποιος άλλος αλγόριθμος όπως ο Hill-Climber.

5. Συμπεράσματα

Στην παρούσα εργασία παρουσιάστηκαν τα Bayesian δίκτυα και τα βασικά χαρακτηριστικά τους. Στη συνέχεια αναλύθηκαν βασικές χρήσεις των δικτύων αυτών και προβλημάτων πάνω τους, όπως είναι η εύρεση δομής του δικτύου και ο συμπερασμός. Κατόπιν ερευνήθηκαν τρόποι για τη λύση του προβλήματος της εύρεσης δομής του δικτύου και έγινε σύγκριση μεταξύ προσεγγιστικών και δυναμικών μεθόδων. Πραγματοποιήθηκε η υλοποίηση ενός αλγορίθμου εύρεσης δομής σε δίκτυα Bayes με ακρίβεια, με σκοπό την εξέταση της αποδοτικότητας ενός τέτοιου αλγορίθμου σε σχέση με τους υπάρχοντες προσεγγιστικούς. Με την εκτέλεση πειραμάτων πάνω στο πεδίο αυτό ανακαλύφθηκε η ανωτερότητα του αλγορίθμου Rebel σε σχέση με τους υπόλοιπους όσον αφορά την πιστότητα των αποτελεσμάτων, αφού οι προσεγγιστικοί αλγόριθμοι έχουν ένα ακόμη επίπεδο αναξιοπιστίας, γεγονός που αντικατοπτρίζεται στα πειράματα. Πάραυτα, οι χρόνοι εκτέλεσης του αλγορίθμου που εξετάστηκε ήταν αρκετά μεγαλύτεροι από τους χρόνους των υπολοίπων κάτι που θα μπορούσε με την κατάλληλη έρευνα και αλλαγές στην υλοποίηση (χωρίς να θυσιαστεί η ορθότητα των αποτελεσμάτων) να βελτιωθεί στο μέλλον.

ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

Ξενόγλωσσος όρος	**Ελληνικός Όρος**
Node	κόμβος
State	κατάσταση
Edge	πλευρά
Posterior	μεταγενέστερη
a-priori	εκ των προτέρων
Parent	Πατέρας
likelihood	πιθανότητα
conditional probability table	πίνακας υπό συνθήκη πιθανοτήτων
Beliefs	πεποιθήσεις
Evidence	ένδειξη
Inference	Συμπερασμός
Learning	Μάθηση
Fully observed variables	Πλήρως γνωστές μεταβλητές
Parameter learning	Εκτίμηση Παραμέτρων Μοντέλου
joint probability distribution	από κοινού κατανομή πιθανοτήτων
probabilistic inference	πιθανολογικός συμπερασμός
Variable elimination	Απαλοιφή μεταβλητών
Clique tree propagation	Διάδοση «δέντρου-κλίκας»
Recursive conditioning	Αναδρομικές ρυθμίσεις
And/or search	ή/και αναζήτηση
Approximate inference algorithm	Αλγόριθμος προσεγγιστικής επαγωγής συμπερασμάτων
loopy belief propagation	αναδραστική διάδοση πίστεως
variational methods	μεταβολικές μέθοδοι
Machine learning	Μηχανική μάθηση
Directed acyclic graph	Κατευθυνόμενο άκυκλο γράφημα
Scoring function	Μέθοδος βαθμολόγησης
Exhaustive search	Εξαντλητική-πλήρης αναζήτηση
Local minima	Τοπικά ελάχιστο
Inductive logic programming	Επαγωγικός λογικός προγραμματισμός
Statistical relational learning	Στατιστική σχεσιακή μάθηση
Parse tree	Συντακτικό δέντρο
Probabilistic graphical models	Πιθανοτικά γραφικά μοντέλα
Branch and bound search	Αναζήτηση διακλάδωσης και οριοθέτησης
Markov random fields	Στοχαστικά πεδία Markov
Inseparability	Μη-διαχωρισμός
Feedback	Ανατροφοδότηση
Terminal state	Τελική κατάσταση
maximum in-degree	Μέγιστη τιμή εισερχομένων ακμών
integrated	ολοκληρωμένο
Causal	Αιτιολογικό

network framework	Σύστημα δικτύου
Ontology driven	με γνώμονα την οντολογία
Vertical integration	Κάθετη ολοκλήρωση
Features	Χαρακτηριστικά
semantic interoperability	Σημασιολογική διαλειτουργικότητα
Markov blanket	Markov κάλυμμα
Toolkit	Εργαλειοθήκη
classification	ταξινόμηση
Regression	Παλινδρόμηση
Clustering	Συσταδοποίηση
Visualization	Οπτικοποίηση
conditional independence assertions	υπό-όρους ισχυρισμούς ανεξαρτησίας

ΑΝΑΦΟΡΕΣ

[1] F. V. Jensen, An Introduction to Bayesian Networks, Springer Verlag, New York, 1996.

[2] Kannan R., Bayesian networks: Application in safety instrumentation and risk reduction, ISA Transactions, 2006.

[3] Rebane, G. and Pearl, J., The Recovery of Causal Poly-trees from Statistical Data, 3rd Workshop on Uncertainty in AI, Seattle, WA, 1987

[4] Friedman Nir, Geiger Dan, Goldszmidt, Moises (November 1997). "Bayesian Network Classifiers"

[5] Nassif Houssam, Kuusisto Finn, Burnside, Elizabeth S, Page David, Shavlik Jude, Santos Costa Vitor, (2013). "Score As You Lift (SAYL): A Statistical Relational Learning Approach to Uplift Modeling"

[6] Gökhan Baklr, Ben Taskar, Thomas Hofmann, Bernhard Schölkopf, Alex Smola and SVN Vishwanathan (2007), Predicting Structured Data, MIT Press.

[7] Cassio P. de Campos, Qiang Ji, Efficient Structure Learning of Bayesian Networks using Constraints, 2011

[8] Loukia Meligkotsidou, Bayes notes, [online], http://eclass.uoa.gr, 2008

[9] Gregory F. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks (research note). Artificial Intelligence, 1990.

[10] Kevin Murphy, YairWeiss, Loopy belief propagation for approximate inference: An empirical study, 1999.

[11] M.J. Wainwright, T.S. Jaakkola, A.S. Willsky. MAP estimation via agreement on trees: message passing and linear programming. IEEE Transactions on Information Theory, 2005.

[12] Ψωϊνός Δ, Ποσοτική Ανάλυση, 2η Έκδοση, Θεσσαλονίκη, Εκδόσεις Ζήτη, 1996.

[13] Mikko Koivisto, Kismat Sood, Exact Bayesian Structure Discovery in Bayesian Networks, Journal of Machine Learning Research 5, 2004.

[14] Prototype implementation of the Knowledge Discovery methods, Information Society Technologies, 2008

[15] Εκπαίδευση ΤΝΔ με το λογισμικό WEKA
http://www.draftpad.gr/2014/02/weka.html

[16] WEKA: ένα ισχυρό εργαλείο στην επιστήμη του Data Mining
http://www.justtech.gr/2015/03/weka-data-mining-tool/

[17] Mikko Koivisto, Advances in Exact Bayesian Structure Discovery in Bayesian Networks, 2006

[18] The University of Waikato, WEKA Manual for Version 3-7-12, 2014