Δίκτυα Bayes: Ένα εργαλείο απόφασης για την αξιολόγηση της πιστοληπτικής ικανότητας

Ζάρδα Ιωάννα
Α.Μ. 09311006

Επιβλέποντες Καθηγητές: Α.Χριστόπουλος
Β.Ζαρίκας

Διαμεταπτυχιακό-Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών με τίτλο:

Μαθηματική Προτυποποίηση στις Σύγχρονες Τεχνολογίες και την Οικονομία

Εθνικό Μετσόβιο Πολυτεχνείο

Κεφάλαιο 1: Εισαγωγή

1.1 Πιστοληπτική Ικανότητα

Ως πιστοληπτική ικανότητα η αξιοπιστία και η φερεγγυότητα ενός προσώπου, φυσικού ή νομικού, στην αποπληρωμή των χρεών του. Η πιστοληπτική ικανότητα αποκαλύπτει σε ένα δανειστή ή επενδυτή την πιθανότητα να μπορέσει ο δανειολήπτης να ανταποκριθεί στις δανειακές του υποχρεώσεις χωρίς τον κίνδυνο πτώχευσης.

1.2 Αξιολόγηση πιστοληπτικής ικανότητας

Ως αξιολόγηση πιστοληπτικής ικανότητας ορίζεται η διαδικασία κατά την οποία αξιολογείται η φερεγγυότητα ενός ατόμου, φυσικού ή νομικού, βάσει κανόνων που ορίζονται τόσο από διεθνείς οργανισμούς αλλά και από τον δανειστή ή τον οίκο αξιολόγησης κατά περίπτωση.

1.3 Δίκτυα Bayes

Ως δίκτυα Bayes ορίζονται πιθανοτικά γραφικά μοντέλα που αναπαριστούν ένα σύνολο τυχαίων μεταβλητών και τις αλληλεξαρτήσεις τους μέσα από έναν κατευθυνόμενο ακυκλικό γράφο (DAG) και χρησιμοποιούνται για την εξαγωγή συμπεράσματος σε συνθήκες αβεβαιότητας.

1.4 Αντικείμενο της Μεταπτυχιακής Εργασίας

Το αντικείμενο της εργασίας αυτής είναι η διερεύνηση της χρήσης των δικτύων Bayes ως εναλλακτικού τρόπου αξιολόγησης της πιστοληπτικής δυνατότητας προσώπων, φυσικών ή νομικών, και τα πλεονεκτήματα που προσφέρουν έναντι άλλων συμβατικών μεθόδων βάσει του μοντέλου παλινδρόμησης Probit.

Κεφάλαιο 2: Γενικές Γνώσεις

2.1 Εισαγωγή

Στο κεφάλαιο αυτό θα αναλυθούν οι παράγοντες που επηρεάζουν την πιστοληπτική ικανότητα, τα μοντέλα τα οποία χρησιμοποιούνται μέχρι τώρα και θα γίνει μια εισαγωγή στα δίκτυα αποφάσεων και ιδιαίτερα στα δίκτυα Bayes και τις βασικές αρχές που τα διέπουν και θα χρησιμοποιηθούν στο επόμενο κεφάλαιο.

2.2 Παράγοντες που επηρεάζουν την πιστοληπτική ικανότητα

Οι παράγοντες που επηρεάζουν την πιστοληπτική ικανότητα ενός δανειολήπτη ,άρα και τη δυνατότητα του να δανείζεται χρήματα από χρηματοπιστωτικά ιδρύματα είναι:

- Το οικονομικό ιστορικό
- Το επιτόκιο αποπληρωμής
- Η διαθεσιμότητα των περιουσιακών στοιχείων
- Η ρευστότητα
- Η παρούσα οικονομική κατάσταση
- Το πιθανό μελλοντικό εισόδημα
- Η αποταμιευτική συμπεριφορά
- Η καταναλωτική συμπεριφορά, καθώς και
- Το ύψος των υποχρεώσεων (δηλαδή αν υπάρχουν άλλα χρέη και δάνεια).

2.3 Δίκτυα Bayes

Εισαγωγή

Τα δίκτυα Bayes είναι γραφικά μοντέλα που συνδέουν ένα σύνολο μεταβλητών με σχέσεις πιθανοτήτων. Τα δίκτυα Bayes έχουν γίνει δημοφιλή για την ικανότητά

τους να ενσωματώνουν την γνώση ειδικών σε ένα έμπειρο σύστημα. Μάλιστα, έχουν αναπτυχθεί μέθοδοι που επιτυγχάνουν τη μάθηση των δικτύων Bayes από δεδομένα που είναι αρκετά αποτελεσματικές σε ορισμένα προβλήματα ανάλυσης δεδομένων.

Παρακάτω ορίζεται ένα δίκτυο Bayes, καθώς και σχετικές μέθοδοι που χρησιμοποιούνται για να αντληθεί και να κωδικοποιηθεί γνώση από τα δεδομένα. Υπάρχουν αρκετοί τρόποι αναπαράστασης για ένα πρόβλημα ανάλυσης δεδομένων, όπως τα δέντρα, τα νευρωνικά δίκτυα, καθώς και αρκετές μέθοδοι για ανάλυση δεδομένων όπως εκτίμηση πυκνότητας και παλινδρόμηση. Οπότε γεννάται το ερώτημα τι προσφέρουν παραπάνω τα δίκτυα Bayes από τις υπόλοιπες μεθόδους.

Πρώτον, τα δίκτυα Bayes μπορούν εύκολα να αντιμετωπίσουν τα ελλιπή σύνολα δεδομένων. Έστω ένα πρόβλημα παλινδρόμησης όπου μεταξύ δύο επεξηγηματικών μεταβλητών δεν υπάρχει συσχετισμός. Οι κλασσικές μέθοδοι μάθησης δεν αντιμετωπίζουν κάποιο πρόβλημα στην περίπτωση που είναι γνωστά όλα τα δεδομένα. Όμως όταν κάποιες παρατηρήσεις δεν είναι γνωστές, τότε οι περισσότερες μέθοδοι δίνουν μια λανθασμένη εκτίμηση γιατί δεν μπορούν να ενσωματώσουν το συσχετισμό μεταξύ των επεξηγηματικών μεταβλητών. Αντίθετα τα δίκτυα Bayes προσφέρουν ένα φυσικό τρόπο να αντιμετωπίζουν τέτοιες εξαρτήσεις.

Επιπλέον, τα δίκτυα Bayes επιτρέπουν να μάθει κάποιος τις αιτιώδεις σχέσεις μεταξύ των μεταβλητών. Αυτό είναι σημαντικό για τουλάχιστον δύο λόγους. Πρώτον, η διαδικασία είναι ιδιαίτερα χρήσιμη όταν προσπαθεί κάποιος να καταλάβει ένα πρόβλημα, όπως κατά τη διάρκεια της επεξηγηματικής ανάλυσης δεδομένων. Επίσης η γνώση των αιτιωδών σχέσεων επιτρέπει να γίνονται προβλέψεις κατά τη διάρκεια παρεμβάσεων. Για παράδειγμα, ένας αναλυτής του μάρκετινγκ μπορεί να θέλει να μάθει αν αξίζει ή όχι να αυξήσει τη διαφήμιση ενός προϊόντος προκειμένου να αυξηθούν οι πωλήσεις του. Για να δώσει μια απάντηση ο αναλυτής θα πρέπει να γνωρίζει αν η διαφήμιση επηρεάζει τις πωλήσεις και σε ποιο βαθμό. Τα δίκτυα Bayes δίνουν απάντηση σε τέτοια

ερωτήματα ακόμη και αν δεν έχει προηγηθεί έρευνα για τα αποτελέσματα της αυξημένης προβολής. Επίσης, τα δίκτυα Bayes σε αντίθεση με τις στατιστικές μεθόδους Bayes διευκολύνουν το συνδυασμό γνωστικής περιοχής και δεδομένων. Όποιος έχει πραγματοποιήσει μια ανάλυση γνωρίζει τη σημασία της πρότερης γνώσης, ειδικά όταν τα δεδομένα είναι σπάνια ή ακριβά. Το γεγονός ότι κάποια συστήματα μπορούν να φτιαχτούν αποκλειστικά από πρότερη γνώση, τονίζει τη σημασία της πρότερης γνώσης. Τα δίκτυα Bayes έχουν αιτιακά χαρακτηριστικά που κάνουν ιδιαίτερα εύκολη την κωδικοποίηση της αιτιακής πρότερης γνώσης. Επιπλέον τα δίκτυα Bayes κωδικοποιούν το μέγεθος της αιτιακής σχέσης με πιθανότητες. Επομένως, πρότερη γνώση και δεδομένα μπορούν να μελετηθούν σε συνδυασμό με μεθόδους της Bayesian στατιστικής.

Τέλος, οι Bayesian μέθοδοι σε συνδυασμό με τα δίκτυα Bayes και άλλων τύπων μοντέλων προσφέρουν μια ικανοποιητική και βασική προσέγγιση ώστε να αποφεύγεται η υπερβολική χρήση δεδομένων. Στα δίκτυα Bayes δεν υπάρχει λόγος να κρατηθούν κάποια από τα διαθέσιμα δεδομένα για εξέταση. Χρησιμοποιώντας τη Bayesian προσέγγιση, τα μοντέλα μπορούν να χρησιμοποιηθούν έτσι ώστε όλα τα διαθέσιμα δεδομένα να χρησιμοποιηθούν για εκπαίδευση.

Η Bayesian προσέγγιση στην πιθανότητα και τη στατιστική

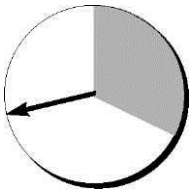
Η κατανόηση των δικτύων Bayes και οι αντίστοιχες μέθοδοι μάθησης επιτυγχάνονται μέσω της κατανόησης της Bayesian προσέγγισης της πιθανότητας και της στατιστικής. Παρακάτω παρουσιάζεται μια εισαγωγή στην Bayesian προσέγγιση για όσους είναι εξοικειωμένοι μόνο με την κλασσική περίπτωση.

Σύμφωνα με την Bayesian προσέγγιση, η πιθανότητα να συμβεί ένα γεγονός είναι η προσωπική πεποίθηση κάποιου. Ενώ η κλασσική πιθανότητα είναι η φυσική συνέπεια του κόσμου, η Bayesian πιθανότητα είναι συνέπεια του ανθρώπου που την ορίζει. Μια βασική διαφορά μεταξύ της κλασσικής και της

Bayesian πιθανότητας είναι ότι για να μετρήσουμε τη Bayesian πιθανότητα δεν χρειάζονται επαναλαμβανόμενες δοκιμές. Για παράδειγμα, υποθέτουμε επαναλαμβανόμενες ρίψεις ενός κύβου ζάχαρης πάνω σε μια υγρή επιφάνεια. Κάθε φορά που ο κύβος πέφτει, οι διαστάσεις του αλλάζουν. Έτσι, ενώ ο κλασσικός στατιστικολόγος ξοδεύει αρκετό χρόνο ώστε να υπολογίσει την πιθανότητα ο κύβος να προσγειωθεί με μια συγκεκριμένη πλευρά προς τα πάνω, ο Bayesian απλά περιορίζει την προσοχή του στην επόμενη ρίψη και ορίζει μια πιθανότητα.

Μια συνήθης κριτική για τον ορισμό της Bayesian πιθανότητας είναι ότι οι πιθανότητες φαίνονται ασαφείς. Προκύπτουν ερωτήματα, όπως γιατί η πεποίθηση κάποιου ικανοποιεί τους κανόνες της πιθανότητας ή πώς θα έπρεπε να μετριούνται οι πιθανότητες. Συγκεκριμένα, έχει νόημα να αντιστοιχούμε μια πιθανότητα που είναι ίση με ένα (μηδέν) σε ένα γεγονός που (δεν) θα συμβεί, αλλά ποια πιθανότητα αντιστοιχούμε σε μια πεποίθηση που δεν είναι στα άκρα; Όλα αυτά τα ερωτήματα έχουν μελετηθεί εκτενώς.

Όσον αφορά το πρώτο ερώτημα ,αρκετοί ερευνητές έχουν προτείνει διάφορα σύνολα κανόνων που θα πρέπει να ικανοποιούν οι πεποιθήσεις. Αποδεικνύεται όμως τελικά ότι κάθε σύνολο αυτών των κανόνων οδηγεί να ικανοποιεί τους κανόνες της πιθανότητας. Παρόλο που κάθε σύνολο αυτών των κανόνων κρίνεται αναγκαίο, το γεγονός ότι διαφορετικά σύνολα οδηγούν στο να ικανοποιούν όλα τους κανόνες της πιθανότητας δίνουν το δικαίωμα να χρησιμοποιηθεί η πιθανότητα για τη μέτρηση της πεποίθησης.



Εικόνα 1: Ο τροχός της πιθανότητας

Όσον αφορά το πώς θα έπρεπε να μετριούνται οι πιθανότητες, η απάντηση δίνεται παρατηρώντας ότι οι άνθρωποι τείνουν να θεωρούν αξιοκρατικό να λένε ότι δύο γεγονότα θεωρούνται εξίσου πιθανά. Έστω ένας τροχός τύχης που είναι

χωρισμένος σε δύο μόνο μέρη(σκιασμένη και μη σκιασμένη περιοχή), όπως στην Εικόνα 1. Υποθέτοντας ότι ο τροχός είναι συμμετρικός (εκτός της σκίασης),θεωρείται εξίσου πιθανό ο τροχός να σταματήσει σε οποιαδήποτε θέση. Λαμβάνοντας υπόψη αυτή την άποψη, καθώς και ότι το άθροισμα των πιθανοτήτων ισούται με ένα, συμπεραίνουμε ότι η πιθανότητα ο τροχός να σταματήσει στην σκιασμένη περιοχή ισούται με το ποσοστό της περιοχής του τροχού που είναι σκιασμένη(στην συγκεκριμένη περίπτωση είναι ίση με 0,3).

Ο τροχός της πιθανότητας προσφέρει έναν τρόπο μέτρησης πιθανοτήτων άλλων γεγονότων. Για παράδειγμα, ποια πιστεύουμε ότι είναι η πιθανότητα να κερδίσει ο Ολυμπιακός το επόμενο πρωτάθλημα; Πρώτα, πρέπει να αναρωτηθούμε αν είναι πιο πιθανό ότι θα κερδίσει ο Ολυμπιακός ή ο τροχός όταν γυρίσει θα σταματήσει στην σκιασμένη περιοχή; Αν πιστεύουμε ότι είναι πιο πιθανό να κερδίσει ο Ολυμπιακός, τότε ας φανταστούμε έναν τροχό με μεγαλύτερη σκιασμένη περιοχή. Αν πιστεύουμε ότι είναι πιο πιθανό ο τροχός να σταματήσει στην σκιασμένη περιοχή, τότε ας φανταστούμε έναν άλλον τροχό που η σκιασμένη περιοχή είναι μικρότερη. Τώρα, επαναλαμβάνουμε τη διαδικασία μέχρι να θεωρήσουμε ισοπίθανα τα γεγονότα να νικήσει ο Ολυμπιακός και ο τροχός να σταματήσει στην σκιασμένη περιοχή. Τότε, η πιθανότητα που πιστεύουμε ότι θα κερδίσει ο Ολυμπιακός είναι ακριβώς ίση με το ποσοστό της σκιασμένης περιοχής του τροχού.

Η παραπάνω μέθοδος που περιγράφηκε σαν ένας τρόπος μέτρησης της πιθανότητας ενός γεγονότος είναι μια από τις πολλές διαθέσιμες στην Διοικητική Επιστήμη ,σε Εργασίες έρευνας και στην Επιστήμη της Ψυχολογίας. Ένα πρόβλημα σε αυτή τη μέθοδο είναι αυτό της αριθμητικής ακρίβειας. Μπορεί κάποιος να πει ότι η δική του ή κάποιου άλλου πιθανότητα είναι ίση με 0,601 και όχι 0,599; Στις περισσότερες περιπτώσεις όχι. Όμως στις περισσότερες περιπτώσεις οι πιθανότητες χρησιμοποιούνται ώστε να ληφθούν αποφάσεις, και οι αποφάσεις αυτές δεν είναι τόσο ευαίσθητες σε μικρές διακυμάνσεις των πιθανοτήτων. Υπάρχουν μέθοδοι που εξετάζουν πότε απαιτείται ή όχι αριθμητική ακρίβεια. Ένα ακόμη πρόβλημα της παραπάνω μεθόδου είναι η λεκτική ακρίβεια.

Για παράδειγμα, πρόσφατες εμπειρίες ή ο τρόπος που μια ερώτηση τίθεται μπορούν να οδηγήσουν σε παραδοχές που δεν εκφράζουν την πραγματική πεποίθηση κάποιου (Tversky and Kahneman, 1974). Μέθοδοι για να αντιμετωπιστεί αυτό το πρόβλημα υπάρχουν στον τομέα της Θεωρίας Ανάλυσης.

Παρακάτω θα αναλυθεί το θέμα της μάθησης με δεδομένα, παραθέτοντας ένα παράδειγμα. Έστω ένας επιχειρηματίας που σκέφτεται να ανοίξει μια επιχείρηση σε μια περιοχή. Σύμφωνα με το επιχειρηματικό του πλάνο, πρέπει να επιτύχει το 25% του αγοραστικού κοινού για να καταστεί η επένδυσή του κερδοφόρα. Η έρευνα μιας εταιρείας δημοσκοπήσεων κατέληξε στο συμπέρασμα ότι από ένα δείγμα 20 καταναλωτών το 25% ενδιαφέρονταν για τις υπηρεσίες της συγκεκριμένης επιχείρησης. Θεωρείται αξιόπιστο όμως το αποτέλεσμα της έρευνας;

Τα δεδομένα που έχει στη διάθεση του ο επιχειρηματίας δεν είναι επαρκή ώστε να θεωρήσει την επένδυσή του κερδοφόρα. Είναι χρήσιμα και άλλα στοιχεία εκτός της δημοσκόπησης και του επιχειρηματικού πλάνου.

Το ιστορικό παρόμοιων επιχειρήσεων είναι πάντα χρήσιμο. Έστω λοιπόν ότι οι νέες επιχειρήσεις καταλαμβάνουν το 25% της αγοράς στο 20% των περιπτώσεων και το 30% στο 40% των περιπτώσεων. Τα στοιχεία δίνονται στον παρακάτω πίνακα.

Ποσοστά νέων επιχειρήσεων που καταλαμβάνουν δεδομένο μερίδιο αγοράς

Μερίδιο Αγοράς	**Ποσοστό επιχειρήσεων**
0.10	0.05
0.15	0.05
0.20	0.20
0.25	0.20
0.30	0.40
0.35	0.10
	Σύνολο = 1.00

Ο επενδυτής πρέπει να δώσει μια απάντηση στο ποια είναι η πιθανότητα η επιχείρησή του να επιτύχει περισσότερο από το 25% του αγοραστικού κοινού ,ώστε σύμφωνα με τα παραπάνω δεδομένα η επιχείρησή του να είναι κερδοφόρα. Άρα πρέπει να βρει την πιθανότητα η επιχείρησή του να ανήκει στο 70% των επιχειρήσεων που κατάφεραν να καταλάβουν το 25% ή περισσότερο του αγοραστικού κοινού. Μια ανάλυση με τη βοήθεια της Bayesian στατιστικής θα δώσει απαντήσεις στο ερώτημά του. Πρώτα όμως θα ορίσουμε το Θεώρημα Bayes.

Θεώρημα του Bayes

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \qquad (1)$$

όπου:

- Το P(A) είναι η εκ των προτέρων πιθανότητα ,δηλαδή τη γνωρίζουμε πριν την εκτέλεση του πειράματος.
- Το P(A|B) είναι η δεσμευμένη πιθανότητα του A δεδομένου του B ή αλλιώς εκ των υστέρων πιθανότητα αφού εξαρτάται από τη δεδομένη τιμή του B.
- Το P(B|A) είναι η δεσμευμένη πιθανότητα του B δεδομένου του A.
- Τέλος το P(B) είναι ανεξάρτητο του A και μπορεί να θεωρηθεί ως ένας παράγοντας εξομάλυνσης.

Εφαρμογή του θεωρήματος Bayes

Οι πιθανότητες στο παράδειγμα του επενδυτή ακολουθούν την Διωνυμική κατανομή .Οπότε ο τύπος που δίνει τις πιθανότητες είναι:

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad (2)$$

Όπου:

- p :η πιθανότητα ενός γεγονότος να συμβαίνει σε κάθε δοκιμή
- x: ο αριθμός των παραπάνω γεγονότων που συμβαίνουν σε κάθε δοκιμή
- n: ο αριθμός των δοκιμών

Σύμφωνα με τον παραπάνω τύπο βρέθηκαν οι παρακάτω πιθανότητες.

Πιθανότητα να βρεθεί ο επενδυτής σε κάθε κατάσταση δεδομένου ότι x=5 και n=20

Γεγονός (Μερίδιο Αγοράς) Pι	**Εκ των προτέρων πιθανότητα Po(Pi)**	**Πιθανότητα κατάστασης P(x=5|pi)**	**Τομή των πιθανοτήτων P(x=5|pi)nPo(pi)**	**Εκ των υστέρων πιθανότητα P (* = 5\ρ ι) P ο (ρ ι)**
				P (χ **= 5)**
0.10	0.05	0.03192	0.001596	0.00959
0.15	0.05	0.10285	0.005142	0.00309
0.20	0.20	0.17456	0.034912	0.20983
0.25	0.20	0.20233	0.040466	0.24321
0.30	0.40	0.17886	0.071544	0.43000
0.35	0.10	0.12720	0.012720	0.07645
Σύνολα	**1.00**	**0.81772**	**0.166381= P(x=5)**	**0.99997**

Ο παράγοντας εξομάλυνσης ισούται με το άθροισμα των τομών όλων των πιθανοτήτων ο οποίος εξαρτάται όπως φαίνεται από το μέγεθος του δείγματος. Αν το δείγμα ήταν μεγαλύτερο ,τότε θα υπήρχαν περισσότερα στοιχεία σε σχέση με αυτά που ήδη ήταν γνωστά. Η τελευταία στήλη δίνει τα αποτελέσματα της εφαρμογής του θεωρήματος Bayes.

Σύμφωνα λοιπόν με την παραπάνω ανάλυση ,βρέθηκε ότι η πιθανότητα η επένδυση να είναι κερδοφόρα είναι 75%. Οπότε τώρα ο επενδυτής έχει ακόμη περισσότερα στοιχεία , ώστε να λάβει την κατάλληλη απόφαση.

Ορισμός ενός δικτύου Bayes

Στο παραπάνω παράδειγμα οι μεταβλητές που χρησιμοποιήθηκαν ήταν λίγες, κάτι όμως που δεν συμβαίνει στα πραγματικά προβλήματα. Τα δίκτυα Bayes είναι η κατάλληλη αναπαράσταση για τέτοιου είδους προβλήματα. Είναι γραφικά μοντέλα που κωδικοποιούν αρκετά καλά την κοινή κατανομή πιθανοτήτων (κλασσική ή Bayesian) για μεγάλο αριθμό μεταβλητών. Παρακάτω ορίζεται ένα δίκτυο Bayes και ο τρόπος που κατασκευάζεται από πρότερη γνώση.

Ένα δίκτυο Bayes ,για ένα σύνολο μεταβλητών X=..., αποτελείται από δύο μέρη:

- ➔ μια δομή δικτύου S που κωδικοποιεί ένα σύνολο ισχυρισμών που είναι υπό όρους ανεξάρτητοι
- ➔ ένα σύνολο P τοπικών πιθανοτήτων που αφορούν κάθε μεταβλητή.

Τα δύο παραπάνω στοιχεία ορίζουν την κοινή κατανομή πιθανότητας για το σύνολο X. Η δομή δικτύου S είναι ένας κατευθυνόμενος ακυκλικός γράφος. Οι κόμβοι στο δίκτυο S είναι αλληλοσυνδεδεμένοι με τις μεταβλητές του συνόλου X. Το σύμβολο X_i συμβολίζει την μεταβλητή αλλά και τον αντίστοιχο κόμβο. Ανάλογα και το σύμβολο Pa συμβολίζει τους γονείς του κόμβου X_i στο δίκτυο S, όπως και τις μεταβλητές που συνδέονται με τους γονείς. Συγκεκριμένα, δεδομένου της δομής S, η κοινή κατανομή πιθανότητας για το σύνολο X δίνεται από τον παρακάτω τύπο.

$$p(x) = \prod_{i=1}^{n} p(x_i|pa_i) \quad (3)$$

Οι τοπικές πιθανότητες κατανομών P είναι οι κατανομές που ικανοποιούν την παραπάνω εξίσωση. Επομένως, το ζεύγος (S,P) κωδικοποιεί την κοινή κατανομή πιθανότητας.

Οι πιθανότητες σε ένα δίκτυο Bayes μπορούν να είναι είτε κλασσικές είτε Bayesian. Όμως όταν τα δίκτυα Bayes φτιάχνονται μόνο έχοντας υπόψη πρότερη

γνώση, τότε οι πιθανότητες θα είναι Bayesian. Όταν φτιάχνονται με βάση κάποια δεδομένα , τότε οι πιθανότητες θα είναι κλασσικές (και οι τιμές τους μπορεί να είναι αβέβαιες).

Παρακάτω δίνεται ένα παράδειγμα με σκοπό να γίνει κατανοητή η διαδικασία με την οποία γίνεται ένα δίκτυο Bayes. Έστω ότι θέλουμε να εξακριβώσουμε αν υπάρχει απάτη παρακολουθώντας τις αγορές μέσω μιας πιστωτικής κάρτας. Πιθανές μεταβλητές είναι οι:

- Απάτη (Α)
- Βενζίνη (Β)
- Κοσμήματα (Κ)
- Ηλικία (Η)
- Φύλο (Φ)

Η μεταβλητή Απάτη συμβολίζει το γεγονός αν η συγκεκριμένη αγορά είναι ή όχι απάτη. Η μεταβλητή Βενζίνη αναπαριστά αν έγινε ή όχι αγορά βενζίνης τις τελευταίες 24 ώρες. Η μεταβλητή Κοσμήματα δηλώνει αν πραγματοποιήθηκε ή όχι αγορά κοσμημάτων τις τελευταίες 24 ώρες. Οι μεταβλητές Ηλικία και Φύλο αφορούν τα στοιχεία του κατόχου της πιστωτικής κάρτας. Οι πιθανότητες για τις διάφορες καταστάσεις αυτών των μεταβλητών δίνονται παρακάτω. Βέβαια σε ένα πραγματικό πρόβλημα ο αριθμός των μεταβλητών θα είναι πολύ μεγαλύτερος. Επίσης θα μπορούσαν να οριστούν με περισσότερες λεπτομέρειες οι μεταβλητές. Για παράδειγμα, η μεταβλητή Ηλικία θα μπορούσε να οριστεί ως συνεχής μεταβλητή.

Το πρώτο βήμα δεν είναι πάντα εμφανές. Θα πρέπει να γίνονται τα παρακάτω βήματα:

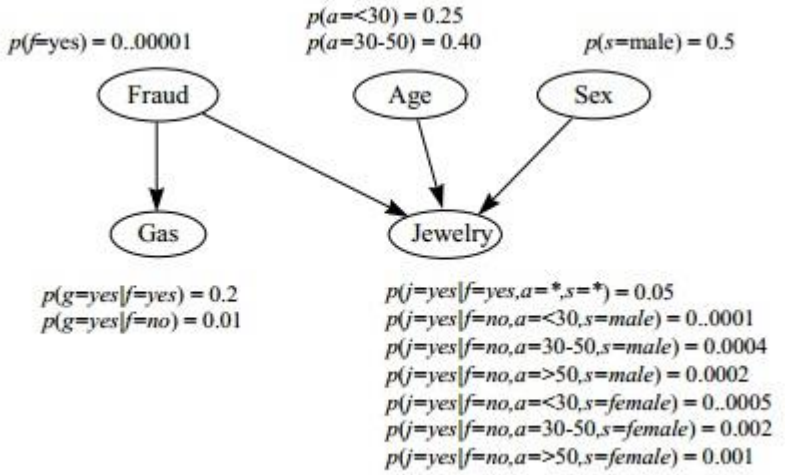
➔ να ορίζονται σωστά οι στόχοι της μοντελοποίησης

➔ να αναγνωρίζονται πολλές πιθανές παρατηρήσεις που μπορεί να είναι σχετικές με το πρόβλημα

➔ να καθορίζεται ποιες από αυτές τις παρατηρήσεις αξίζουν να συμπεριληφθούν στο μοντέλο

➔ να αντιστοιχίζονται όλες οι παρατηρήσεις σε μεταβλητές

Οι δυσκολίες που αντιμετωπίζονται σε αυτό το βήμα δεν είναι θέμα μόνο στα δίκτυα Bayes, αλλά στις περισσότερες μεθόδους. Παρόλο που δεν υπάρχει τέλεια λύση, όσοι ασχολούνται με τη θεωρία ανάλυσης μπορούν να προσφέρουν κάποια καθοδήγηση.



Εικόνα 2 : Ένα δίκτυο Bayes για την εξακρίβωση απάτης παρακολουθώντας τις αγορές μέσω μιας πιστωτικής κάρτας. Τα βέλη ξεκινούν από την αιτία και καταλήγουν στο αποτέλεσμα. Οι πιθανότητες που σχετίζονται με κάθε κόμβο εμφανίζονται δίπλα στον κόμβο. Ο αστερίσκος * σημαίνει ότι η μεταβλητή μπορεί να βρίσκεται σε τυχαία κατάσταση.

Το επόμενο βήμα είναι η κατασκευή ενός κατευθυνόμενου ακυκλικού γράφου που κωδικοποιεί πεποιθήσεις που είναι υπό όρους ανεξάρτητες. Ένας τρόπος για να το επιτύχουμε αυτό βασίζεται στις παρακάτω παρατηρήσεις. Σύμφωνα με τον κανόνα της αλυσίδας της πιθανότητας έχουμε:

$$p(x) = \prod_{i=1}^{n} p(x_i | x_1, \ldots, x_{i-1}) \quad (4)$$

)

Για κάθε X_i υπάρχει ένα υποσύνολο $\Pi_i \subseteq \{X_1,\ldots,X_{i-1}\}$ τέτοιο ώστε τα X_i και το $\{X_1,\ldots,X_{i-1}\} \setminus \Pi_i$ είναι υπό όρους ανεξάρτητα δεδομένου του Π_i. Τότε για κάθε x ,έχουμε:

$$p(x_i|x_1,\ldots,x_{i-1}) = p(x_i|\pi_i) \quad (5)$$

Συνδυάζοντας τις σχέσεις 4 και 5, έχουμε:

$$p(x) = \prod_{i=1}^{n} p(x_i|\pi_i) \quad (6)$$

Συγκρίνοντας τις σχέσεις 3 και 6 ,παρατηρούμε ότι τα σύνολα των μεταβλητών (Π_i ,..., Π_n) αντιστοιχούν στα σύνολα που αποκαλούμαι στα δίκτυα Bayes γονείς (Pa_1,...,Pa_n), τα οποία δικαιολογούν τα βέλη στη δομή δικτύου S.

Επομένως, τα βήματα για να καθορίσουμε τη δομή ενός δικτύου Bayes είναι:

- ➔ βάζουμε σε σειρά τις μεταβλητές με όποιον τρόπο αποφασίσουμε και
- ➔ καθορίζουμε το σύνολο των μεταβλητών που ικανοποιούν την εξίσωση 3 για i=1,...,n.

Στο δικό μας παράδειγμα ,αν χρησιμοποιήσουμε τη σειρά μεταβλητών (A,H,Φ,B,K), τότε οι υπό όρους ανεξαρτησίες θα είναι:

$$p(a|f)=p(a)$$
$$p(s|f,a)=p(s)$$
$$p(g|f,a,s)=p(g|f)$$
$$p(j|f,a,s,g)=p(j|f,a,s) \quad (7)$$

Έτσι, το δίκτυο απέκτησε τη δομή που φαίνεται στην Εικόνα 2.

Η προσέγγιση αυτή έχει ένα σοβαρό μειονέκτημα. Αν δεν επιλέξουμε προσεκτικά τη σειρά των μεταβλητών, η δομή που θα αποκτήσει το δίκτυο μπορεί να αποτύχει να εμφανίσει αρκετές υπό όρους ανεξαρτησίες μεταξύ των μεταβλητών. Για παράδειγμα, αν για το παράδειγμά μας επιλέξουμε να κατατάξουμε τις μεταβλητές ως (K,B,Φ,H,A), τότε το δίκτυο που προκύπτει είναι πλήρως συνδεδεμένο. Η χειρότερη όμως περίπτωση είναι να πρέπει να

ελέγξουμε n! σειρές μεταβλητών, ώστε να βρούμε την καλύτερη. Ευτυχώς υπάρχει άλλη μια τεχνική για την κατασκευή δικτύων Bayes χωρίς να απαιτείται ιεράρχηση. Η τεχνική βασίζεται σε δύο παρατηρήσεις:

➔ οι άνθρωποι τις περισσότερες φορές μπορούν να επιβεβαιώσουν τις αιτιώδεις σχέσεις μεταξύ των μεταβλητών και

➔ οι αιτιώδεις σχέσεις αντιστοιχούν σε ισχυρισμούς που είναι υπό όρους εξαρτημένες.

Συγκεκριμένα, για την κατασκευή ενός δικτύου Bayes για ένα συγκεκριμένο σύνολο μεταβλητών, σχεδιάζουμε βέλη από την αιτία στο άμεσο αποτέλεσμα. Στις περισσότερες περιπτώσεις, αυτή η τεχνική οδηγεί σε μια δομή που ικανοποιεί τον ορισμό της Εξίσωσης 3. Για παράδειγμα, δεδομένου ότι η Απάτη είναι μια μεταβλητή που επηρεάζει άμεσα τη μεταβλητή Βενζίνη, καθώς και οι μεταβλητές Απάτη, Ηλικία και Φύλο επηρεάζουν άμεσα τη μεταβλητή Κόσμημα, αποκτούμε τη δομή που φαίνεται στην Εικόνα 2. Οι αιτιώδεις σχέσεις σε ένα δίκτυο Bayes επηρεάζουν σημαντικά την επιτυχία ενός δικτύου Bayes ως αναπαράσταση ενός έμπειρου συστήματος (Hecherman et al., 1995a).

Το τελευταίο βήμα στην κατασκευή του δικτύου είναι να υπολογίσουμε τις τοπικές πιθανότητες $p(x_i|pa_i)$. Στο δικό μας παράδειγμα, όλες αυτές οι πιθανότητες έχουν υπολογιστεί όπως φαίνεται και στην Εικόνα 2.

Πρέπει να έχουμε υπόψη, πως παρότι παρουσιάσαμε τα βήματα της κατασκευής ενός δικτύου με συγκεκριμένη σειρά στην πράξη πολλές φορές αλλάζουν σειρά. Για παράδειγμα, απόψεις σχετικά με τις υπό όρους ανεξαρτησίες και/ή η αιτία και το αποτέλεσμα μπορούν να επηρεάσουν τη διατύπωση του προβλήματος. Επίσης, αξιολογώντας κάποιες πιθανότητες μπορεί να κριθεί σκόπιμο να αλλάξει η δομή του δικτύου.

Τα συμπεράσματα που εξάγουμε από ένα δίκτυο Bayes

Όταν κατασκευάσουμε το δίκτυο (είτε από πρότερη γνώση είτε από δεδομένα ή συνδυασμό και των δύο), μετά συνήθως χρειαζόμαστε να υπολογίσουμε διάφορες πιθανότητες που μας ενδιαφέρουν από το μοντέλο. Για παράδειγμα,

στο δικό μας παράδειγμα της Απάτης, θέλουμε να μάθουμε την πιθανότητα να υπάρχει Απάτη δεδομένου ότι γνωρίζουμε τις παρατηρήσεις των υπόλοιπων μεταβλητών. Η συγκεκριμένη πιθανότητα δεν βρίσκεται ήδη στο δίκτυο, οπότε πρέπει να υπολογιστεί με τη βοήθεια αυτού. Παρακάτω περιγράφεται πώς υπολογίζουμε τέτοιες πιθανότητες στα δίκτυα Bayes.

Αφού ένα δίκτυο Bayes για μια μεταβλητή X καθορίζει την κοινή κατανομή πιθανότητας για τη X, τότε χρησιμοποιούμε το δίκτυο Bayes για να υπολογίσουμε οποιαδήποτε πιθανότητα μας ενδιαφέρει. Για παράδειγμα, σύμφωνα με το δίκτυο της Εικόνας 2, η πιθανότητα της Απάτης δεδομένου ότι γνωρίζουμε τις παρατηρήσεις των άλλων μεταβλητών υπολογίζεται όπως φαίνεται παρακάτω:

$$p(f|\ a,s,g,j) = \frac{p(f,a,s,g,j)}{p(\ a,s,g,j)} = \frac{p(f,a,s,g,j)}{\sum_{f'} p(f',a,s,g,j)} \quad (8)$$

Όταν όμως ασχολούμαστε με προβλήματα με πολλές μεταβλητές, τότε αυτή η προσέγγιση δεν είναι πρακτική. Ευτυχώς, τουλάχιστον όταν όλες οι μεταβλητές είναι διακριτές, μπορούμε να εκμεταλλευτούμε τις υπό όρους ανεξαρτησίες που είναι κωδικοποιημένες στο δίκτυο Bayes για να κάνουμε τους υπολογισμούς μας περισσότερο αποτελεσματικούς. Στο παράδειγμά μας , γνωρίζοντας τις υπό όρους ανεξαρτησίες της σχέσης 7, η εξίσωση 8 γίνεται:

$$p(f|\ a,s,g,j) = \frac{p(f)p(a)p(s)p(g|f)p(j|f,a,s)}{\sum_{f'} p(f')p(a)p(s)p(g|f')p(j|f',a,s)} =$$

$$= \frac{p(f)p(g|f)p(j|f,a,s)}{\sum_{f'} p(f')p(g|f')p(j|f',a,s)} \quad (9)$$

Παρόλο που χρησιμοποιούμε τις υπό όρους ανεξαρτησίες για να απλοποιήσουμε τις πιθανότητες, ακριβή συμπεράσματα σε ένα αφηρημένο δίκτυο Bayes για διακριτές μεταβλητές είναι εξαιρετικά δύσκολο (Cooper, 1990). Η πηγή του προβλήματος είναι οι μη κατευθυνόμενοι κύκλοι σε ένα δίκτυο Bayes, δηλαδή οι κύκλοι που υπάρχουν στο δίκτυο αν αγνοήσουμε την κατεύθυνση των βελών. Για παράδειγμα, αν στο δίκτυο της Εικόνας 2 προσθέσουμε ένα βέλος

από τη μεταβλητή Ηλικία προς τη μεταβλητή Βενζίνη, τότε στο δίκτυο θα έχουμε έναν μη κατευθυνόμενο κύκλο: Α-Β-Η-Κ-Α). Όταν σε ένα δίκτυο Bayes περιέχονται πολλοί μη κατευθυνόμενοι κύκλοι, τότε είναι δύσκολο να εξάγουμε συμπεράσματα. Ευτυχώς σε αρκετές περιπτώσεις η δομή των δικτύων είναι αρκετά απλή (ή τουλάχιστον μπορεί να απλοποιηθεί ικανοποιητικά χωρίς αυτό να επηρεάζει την ακρίβεια) ώστε να μπορούμε να βγάλουμε ικανοποιητικά συμπεράσματα.

Η εκμάθηση στα δίκτυα Bayes

Σε πολλές περιπτώσεις το δίκτυο Bayes είναι άγνωστο, οπότε πρέπει να το κατασκευάσουμε από τα δεδομένα που έχουμε διαθέσιμα. Αυτό το πρόβλημα είναι γνωστό ως πρόβλημα εκμάθησης στα δίκτυα Bayes Το πρόβλημα μπορεί να διατυπωθεί αλλιώς και ως εξής: Αν γνωρίζεις δεδομένα και πρότερη πληροφορία (δηλαδή πληροφορίες από ειδικούς, τις αιτιώδεις σχέσεις μεταξύ των μεταβλητών), τότε καθόρισε τη δομή του δικτύου και τις παραμέτρους της κοινής κατανομής πιθανότητας στο δίκτυο Bayes.

Ο καθορισμός της δομής ενός δικτύου Bayes είναι δυσκολότερο πρόβλημα από τον καθορισμό των παραμέτρων που θα είναι στο δίκτυο Bayes. Ένα ακόμη πρόβλημα που αντιμετωπίζουμε είναι όταν δεν γνωρίζουμε όλες τις παρατηρήσεις ,είτε αυτό σημαίνει ότι κάποιοι κόμβοι είναι κρυμμένοι είτε ότι κάποια δεδομένα λείπουν. Γενικά, τέσσερις περιπτώσεις εκμάθησης ενός δικτύου Bayes συναντάμε. Η κάθε μια περίπτωση αντιμετωπίζεται χρησιμοποιώντας διαφορετικές μέθοδοι εκμάθησης κάθε φορά ,όπως φαίνεται και στον παρακάτω πίνακα.

Μέθοδοι εκμάθησης ανάλογα με αυτά που ήδη ξέρουμε για το πρόβλημα (Murphy and Mian (1999))

Structure	Observability	Method
Known	full	Maximum-likelihood estimation
Known	partial	EM (or gradient ascent), MCMC
Unknown	full	Search through model space
Unknown	partial	Structural EM + search through model space

- Στην πρώτη και πιο εύκολη περίπτωση, στόχος μας είναι να υπολογίσουμε τις τιμές των παραμέτρων του δικτύου Bayes που μεγιστοποιούν την πιθανότητα (log) του συνόλου δεδομένων που χρησιμοποιείται για την εκμάθηση. Αυτό το σύνολο δεδομένων περιέχει m περιπτώσεις οι οποίες συχνά υποθέτουμε ότι είναι ανεξάρτητες μεταξύ τους. Έστω ότι το σύνολο δεδομένων που χρησιμοποιείται για την εκμάθηση είναι το $E=\{x_i,\ldots,x_m\}$, όπου $x_l=(x_{l1},\ldots,x_{ln})^T$, και το σύνολο των παραμέτρων είναι το $\Theta=(\theta_1,\ldots,\theta_n)$, όπου θ_i είναι το διάνυσμα των παραμέτρων για τη δεσμευμένη πιθανότητα της μεταβλητής X_i (που αναπαρίσταται με έναν κόμβο στο γράφημα). Τότε η log-πιθανότητα του συνόλου δεδομένων που χρησιμοποιείται για την εκμάθηση είναι ένα άθροισμα όρων, για κάθε κόμβο ξεχωριστά:

$\log L(\Theta|\Sigma)=\Sigma\Sigma \log P(x_{li}|\pi_i,\theta_i)$

Η log-πιθανότητα της απόδοσης των αποφάσεων που παίρνουμε όταν υπάρχει αβεβαιότητα *αποσυντίθεται* σύμφωνα με τη δομή του δικτύου. Για παράδειγμα, κάποιος μπορεί να μεγιστοποιήσει την συνεισφορά της log-πιθανότητας κάθε κόμβου ξεχωριστά. Μια άλλη εναλλακτική είναι να ορίσουμε μια προγενέστερη συνάρτηση πυκνότητας πιθανότητας για κάθε διάνυσμα παραμέτρου και να χρησιμοποιήσουμε τα δεδομένα που χρησιμοποιούμε για την εκμάθηση για να υπολογίσουμε τη μεταγενέστερη κατανομή των παραμέτρων και τις εκτιμήσεις Bayes.

- Γενικά, οι υπόλοιπες περιπτώσεις εκμάθησης είναι υπολογιστικά δύσκολες. Στη δεύτερη περίπτωση όπου γνωρίζουμε τη δομή του δικτύου αλλά μας λείπουν κάποια δεδομένα, χρησιμοποιούμε τον αλγόριθμο ΕΜ (μεγιστοποίηση της προσδοκίας) για να βρούμε μια τοπικά βέλτιστη μέγιστη πιθανότητα εκτίμησης των παραμέτρων. Μπορεί να χρησιμοποιήσουμε επίσης τη μέθοδο MCMC για να εκτιμήσουμε τις παραμέτρους ενός δικτύου Bayes.
- Στην τρίτη περίπτωση στόχος μας είναι να βρούμε εκείνο το κατευθυνόμενο ακυκλικό γράφημα που ικανοποιεί καλύτερα τα δεδομένα μας. Αυτή η περίπτωση είναι αρκετά δύσκολη, αφού τα κατευθυνόμενα ακυκλικά γραφήματα που πρέπει να ελέγξουμε είναι πάρα πολλά. Ένας τρόπος είναι να υποθέσουμε ότι όλες οι μεταβλητές είναι υπό όρους ανεξάρτητες δεδομένης μιας κλάσης. Αυτή η περίπτωση αναπαρίσταται στο γράφημα έχοντας όλες οι μεταβλητές τον ίδιο κόμβο γονιό. Παρόλο που είναι ένα πολύ απλό δίκτυο Bayes, δίνει πολύ καλά αποτελέσματα σε ορισμένα προβλήματα.
- Στην τέταρτη και τελευταία περίπτωση δεν γνωρίζουμε πλήρως ούτε τη δομή του δικτύου ούτε τα δεδομένα. Οι Murphy και Mian (1999) έχουν συνοψίσει μια μέθοδο που προτείνει ο Friedman για την συγκεκριμένη περίπτωση όπως περιγράφεται παρακάτω.

1. Προσθέτουμε έναν νέο κόμβο στο δίκτυο που αναπαριστά μια κρυμμένη μεταβλητή
2. Για τους κόμβους που έχουμε, βρίσκουμε το καλύτερο δυνατό δίκτυο που τους ενώνει

3. Συνεχίζουμε την ίδια διαδικασία όσο το δίκτυο μας κάθε φορά βελτιώνεται

Κεφάλαιο 3: Σχεδιασμός δικτύου Bayes για δάνεια μικρών επιχειρήσεων με το Genie

3.1 Εισαγωγή

Στο κεφάλαιο αυτό παρουσιάζεται το μοντέλο αξιολόγησης δανείων μικρών επιχειρήσεων, το οποίο με την βοήθεια του προγράμματος Genie θα μετατραπεί σε δίκτυο Bayes.

3.2 Μοντέλο αξιολόγησης δανείων μικρών επιχειρήσεων

Πολλές τράπεζες έχουν αυξήσει το ενδιαφέρον τους να αναπτύξουν τις σχέσεις τους με μικρές επιχειρήσεις, παρόλο που η δανειοδότηση αυτών θεωρείται ότι έχει υψηλό ενδεχόμενο ρίσκο. Ένας από τους λόγους που αυξάνουν το ρίσκο είναι η ασυμμετρική πληροφόρηση, το γεγονός δηλαδή ότι η τράπεζα μπορεί να έχει λιγότερες πληροφορίες για την κατάσταση μιας επιχείρησης από ότι η ίδια η επιχείρηση κάτι που θα μπορούσε να εκμεταλλευτεί προς όφελός της. Μια επιχείρηση συνήθως είναι πολύ καλύτερα πληροφορημένη για την ικανότητά της, καθώς και την προθυμία της να αποπληρώσει ένα δάνειο , Επίσης η οικονομική δραστηριότητα μιας μικρής επιχείρησης δεν είναι τόσο σταθερή στο χρόνο όσο σε μεγαλύτερες επιχειρήσεις.

Στο εθνικό πανεπιστήμιο Chengchi της Ταιβάν πραγματοποιήθηκε μια μελέτη, με σκοπό να αναπτυχθεί ένα μοντέλο βαθμολόγησης της πιστοληπτικής ικανότητας μικρών επιχειρήσεων. Στόχος τους ήταν να μειώσουν τις επιπτώσεις της ασυμμετρικής πληροφόρησης, δηλαδή την «δυσμενή επιλογή» και τον «ηθικό κίνδυνο». Αρχικά συνέλλεξαν πληροφορίες που αφορούσαν τα βασικά χαρακτηριστικά μιας μικρής επιχείρησης. Επίσης χρησιμοποίησαν πληροφορίες σχετικά με τη σχέση που είχαν ήδη αναπτύξει οι μικρές επιχειρήσεις με τράπεζες σε προηγούμενες συναλλαγές τους, καθώς και το ιστορικό των ιδιοκτητών των μικρών επιχειρήσεων με τις τράπεζες. Επιπλέον, έπρεπε να μετριάσουν την ευαισθησία της οικονομικής τάσης της δανειοδότησης μικρών επιχειρήσεων. Για να το επιτύχουν αυτό χρησιμοποίησαν παράγοντες που επηρεάζουν τη βιομηχανία, όπως οικονομικούς κύκλους και μακροοικονομικούς παράγοντες.

Ανέπτυξαν ένα μοντέλο βαθμολόγησης πιστοληπτικής ικανότητας χρησιμοποιώντας το μοντέλο της Probit παλινδρόμησης. Ως εξαρτημένη μεταβλητή y όρισαν τη μεταβλητή που εκτιμάει τον πιστωτικό κίνδυνο και την πιθανότητα κινδύνου αθέτησης. Η εξαρτημένη μεταβλητή θα ισούται με 1, όταν η μικρή επιχείρηση αποτύχει να αποπληρώσει το δάνειο, ενώ ο βασικός ιδιοκτήτης έχει πάρει το δάνειο και με 0 όταν δεν αποτύχει. Η μεταβλητή y εκφράζεται ως μια γραμμική συνάρτηση ορισμένων επεξηγηματικών μεταβλητών και ενός όρου v που εκφράζει το σφάλμα. Οι επεξηγηματικές μεταβλητές που περιγράφουν το πιστωτικό ιστορικό του δανειολήπτη συμβολίζονται με x,k,m,z. Επομένως, το μοντέλο που χρησιμοποίησαν δίνεται από την σχέση y=a+bx+ck+dm+ez+v , όπου τα a,b,c,d και e είναι οι αντίστοιχοι σταθεροί συντελεστές.

- Η μεταβλητή x περιέχει πληροφορίες που αφορούν την ταυτότητα της επιχείρησης. Η ταυτότητα μιας επιχείρησης καθορίζεται από πληροφορίες όπως

✓το όνομα της επιχείρησης,

✓την ηλικία (τα χρόνια ύπαρξής της),

✓τα έσοδα από εισαγωγές τα τελευταία τρία χρόνια ,

✓τα έσοδα από εξαγωγές τα τελευταία τρία χρόνια ,

✓η αποδοτικότητα της επιχείρησης ,καθώς και

✓την οικονομική δυσκολία που ίσως αντιμετωπίζει.

- Η μεταβλητή k περιγράφει το πιστωτικό ιστορικό του κύριου ιδιοκτήτη της επιχείρησης. Το πιστωτικό ιστορικό του κύριου ιδιοκτήτη περιγράφεται από πληροφορίες σχετικά με λογαριασμούς πιστωτικών καρτών και προσωπικών δανείων που έχει. Τέτοιες πληροφορίες είναι

✓ο αριθμός των λογαριασμών,

✓το υπόλοιπο ενός δανείου,

✓το ιστορικό πληρωμών ,καθώς και

✓το ιστορικό αθέτησης αποπληρωμής ενός δανείου .

Επίσης η μεταβλητή k δηλώνει αν ο κύριος ιδιοκτήτης της επιχείρησης είναι ιδιοκτήτης και κάποιας άλλης επιχείρησης.

- Η μεταβλητή m περιλαμβάνει μεταβλητές που συνδέονται με το επιχειρηματικό πιστωτικό ιστορικό της επιχείρησης.

✓Ο αριθμός των τραπεζών από τις οποίες μια επιχείρηση έχει λάβει κάποιο δάνειο,

✓ο αριθμός των τραπεζών που έχουν υψηλό κύρος με τις οποίες συνεργάζεται ,

✓η συχνότητα των καταχωρήσεων στο πιστωτικό αρχείο (τους τελευταίους 3 μήνες, τους τελευταίους 6 μήνες και τα τελευταία 3 χρόνια),

✓τα βραχυπρόθεσμα και μακροπρόθεσμα δάνεια ,

✓τα δάνεια με υποθήκη,

✓οι εγγυήσεις ,

✓οι καταθέσεις,

✓εξάρτηση από τραπεζικά δάνεια και

✓το ιστορικό αθέτησης αποπληρωμής των δανείων σε όλες τις τράπεζες.

- Η μεταβλητή z αναπαριστά τις επιπτώσεις της βιομηχανίας. Επίσης η μεταβλητή z χρησιμοποιεί βασικούς μακροοικονομικούς δείκτες για να εκτιμήσει την οικονομική κατάσταση των μικρών επιχειρήσεων κατά τη χρονική περίοδο του δανεισμού.

- Η μεταβλητή ν,σύμφωνα με οικονομετρική ανάλυση που έγινε, είναι μια ανεξάρτητη μεταβλητή που ακολουθεί την κανονική κατανομή με μέση τιμή 0.

Τα δεδομένα που χρησιμοποιήθηκαν στην έρευνα, αφορούν 41,000 μικρές επιχειρήσεις, από τις οποίες οι 6,000 απέτυχαν να αποπληρώσουν τα δάνειά τους και οι υπόλοιπες 35,000 εκπλήρωσαν επιτυχώς τις υποχρεώσεις τους. Στη συγκεκριμένη έρευνα, ως μικρή επιχείρηση ορίζεται κάθε επιχείρηση με κεφαλαιοποίηση μικρότερη των 200 εκατομμυρίων NT$.Χρησιμοποιήθηκαν δεδομένα από την περίοδο 1996 έως 1999 για τον υπολογισμό του μοντέλου και δεδομένα από την περίοδο του 2000 για την επιβεβαίωση του μοντέλου. Η συλλογή των δειγμάτων βασίστηκε σε δύο πηγές. Η βασική πηγή των δεδομένων είναι το Joint Credit information Center της Ταιβάν για την περίοδο 1996-2000.Τα δεδομένα περιέχουν πληροφορίες για τα χαρακτηριστικά της επιχείρησης, την οικονομική κατάσταση της επιχείρησης όπως και δεδομένα που αφορούν τη

σχέση της επιχείρησης με την τράπεζα. Τα παραπάνω δεδομένα ,καθώς και το πιστωτικό ιστορικό του βασικού ιδιοκτήτη και το πιστωτικό ιστορικό της επιχείρησης επιτρέπουν την εκτίμηση της ασυμμετρικής πληροφόρησης και τα δυναμικά χαρακτηριστικά της δανειοδότησης των μικρών επιχειρήσεων. Επίσης χρησιμοποιήθηκαν δεδομένα από το Taiwan Economic Journal για την περίοδο 1996-2000 που έδωσαν πληροφορίες σχετικά με την ταξινόμηση των βιομηχανιών ,τους κλαδικούς δείκτες και βασικούς μακροοικονομικούς δείκτες.

Με βάση τα παραπάνω δεδομένα, εφαρμόστηκε η Probit παλινδρόμηση διαδοχικά μέχρι να βρεθεί το καταλληλότερο μοντέλο προσαρμογής για την περίοδο που εξετάζεται. Σύμφωνα με την ανάλυση που έγινε το μοντέλο που επιλέχθηκε προβλέπει το 80% των περιπτώσεων που θα αποτύχουν να αποπληρώσουν το δάνειο. Οι επεξηγηματικές μεταβλητές που επιλέχθηκαν στο μοντέλο είναι :

✓η ηλικία,

✓η οικονομική δυσκολία,

✓τα προσωπικά δάνεια,

✓αν ο κύριος ιδιοκτήτης είναι ιδιοκτήτης και κάποιας άλλης επιχείρησης.

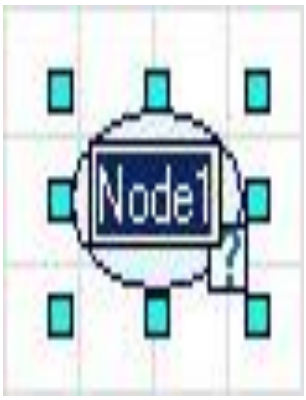
3.3 Λειτουργία του προγράμματος σχεδιασμού δικτύων Bayes Genie

Το Genie είναι ένα πρόγραμμα σχεδιασμού γραφικών μοντέλων αποφάσεων. Αναπτύχθηκε από το Εργαστήριο Συστημάτων Απόφασης του Πανεπιστημίου του Pittsburgh. Έχει πάρει το όνομά του από την αγγλική ορολογία Graphical Network Interface, δηλαδή Διεπαφή Γραφικών Δικτύων.

Παρακάτω θα περιγραφεί ο τρόπος με τον οποίον γίνεται ένας κόμβος και πως δίνονται ιδιότητες σε αυτόν, πως προσδιορίζεται η εξαρτήση με πιθανότητες ανάμεσα στους κόμβους και πως ενημερώνεται το μοντέλο και πως φαίνονται τα αποτελέσματα μετά από την παρατήρηση συγκεκριμένων κόμβων.

- Δημιουργία κόμβου

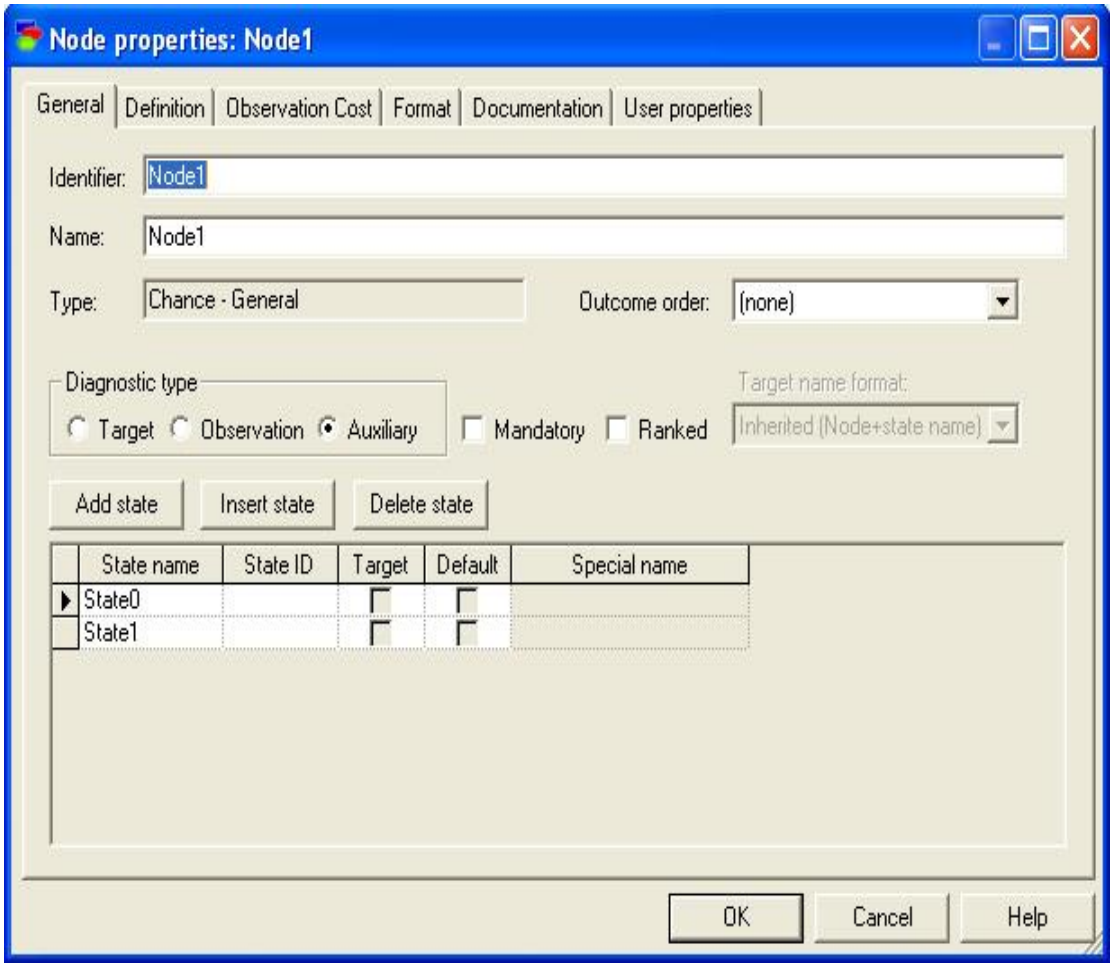
Για την δημιουργία κόμβου πιθανότητας από την εργαλειοθήκη επιλέγεται το εικονίδιο [Chance]. Ο κέρσορας αλλάζει σε και πατώντας στην οθόνη του Genie δημιουργείται ο παρακάτω κόμβος:



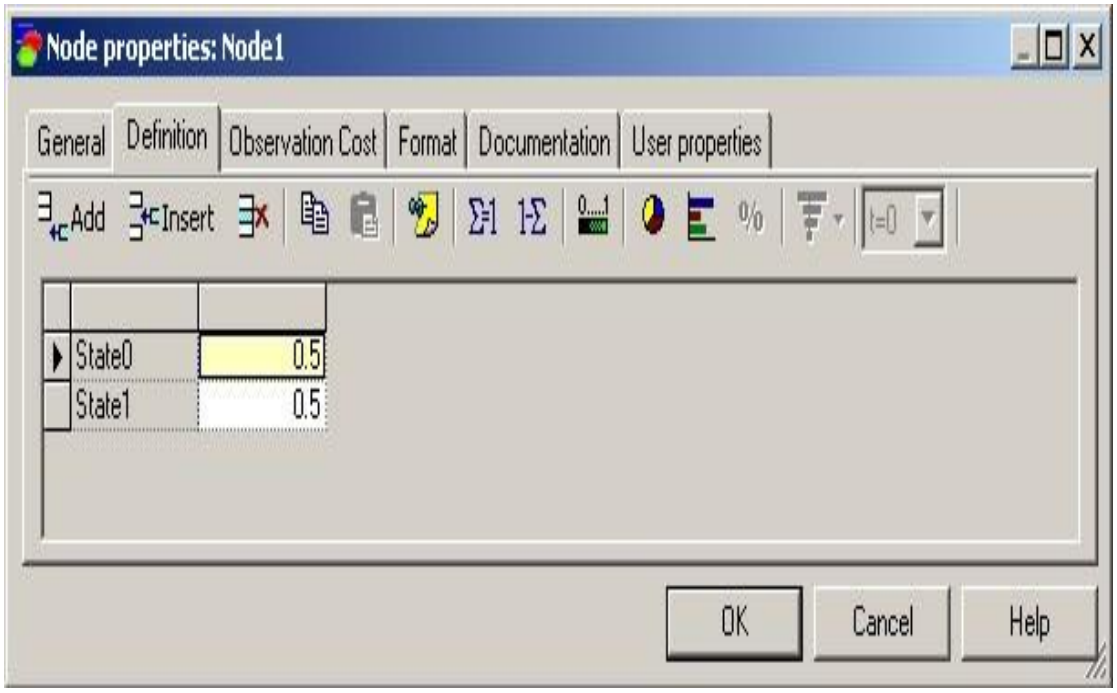
Τα μικρά τετράγωνα γύρω από αυτόν συμβολίζουν ότι ο κόμβος είναι επιλεγμένος, αλλά επίσης ότι το μέγεθός του στην οθόνη μπορεί να αυξομειωθεί. Το όνομα του κόμβου (node1) επιλέγεται αυτόματα από το πρόγραμμα αλλά μπορεί να αλλάξει. Το Genie συσχετίζει δύο ταμπέλες με τον κάθε κόμβο, το όνομα και το αναγνωριστικό του. Το όνομα είναι μια σειρά χαρακτήρων για τον χρήστη και το αναγνωριστικό αφορά το πρόγραμμα.

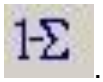
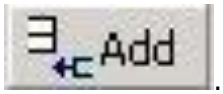
- Προσδιορισμός στοιχείων του κόμβου

Κάνοντας δεξί κλικ στον κόμβο που δημιουργήθηκε πριν, το Genie ανοίγει το παρακάτω παράθυρο με επιλογές:



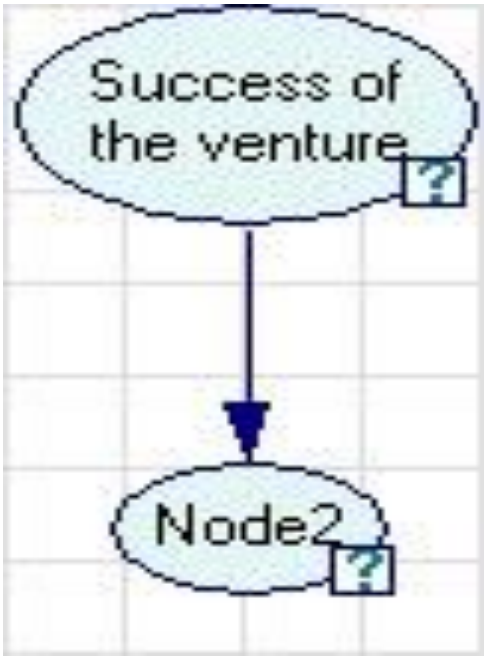
Το παράθυρο αυτό χρησιμοποιείται για να ρυθμιστούν οι ιδιότητες του κόμβου. Στην καρτέλα Definition μπορούν να οριστούν τα αποτελέσματα αυτής της μεταβλητής (κόμβου) ανάλογα με την πιθανότητα τους.



Εδώ μπορούν να αλλάξουν τα ονόματα των αποτελεσμάτων, αλλά και να υπολογιστεί αυτόματα η πιθανότητα από το δεύτερο αποτέλεσμα πατώντας το  . Τέλος, μπορούν εδώ να προστεθούν και άλλες καταστάσεις πατώντας το  .

- Σχέσεις μεταξύ κόμβων

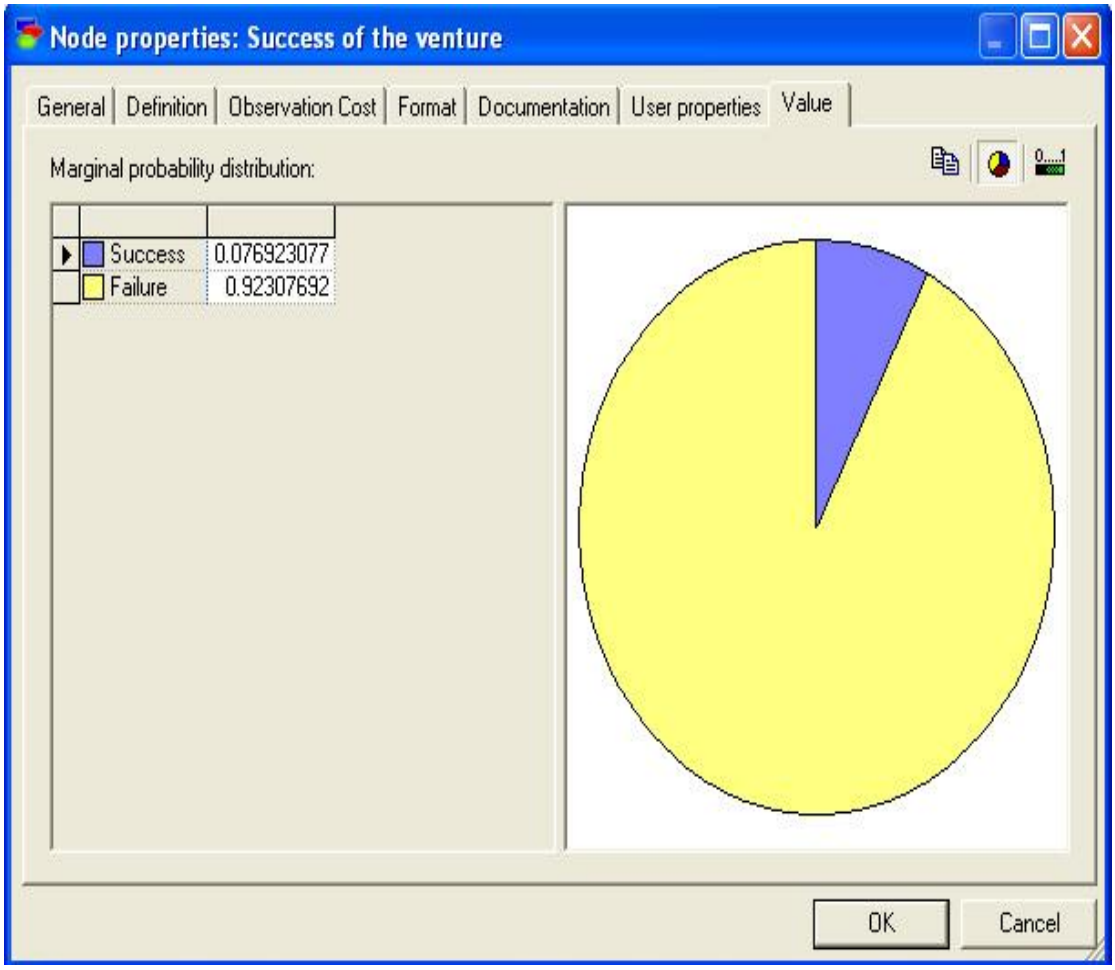
  Για να αναπαρασταθεί το γεγονός ότι δύο κόμβοι σχετίζονται μεταξύ τους, δημιουργείται ένα βέλος επιρροής μεταξύ τους πατώντας το  . Με σύρσιμο του κέρσορα από τον ένα στον άλλον κόμβο σχεδιάζεται το βέλος, που φαίνεται παρακάτω:



Το βέλος ανάμεσα στους δύο κόμβους σημαίνει ότι το αποτέλεσμα του πρώτου θα αποφέρει διαφορά για την διασπορά της πιθανότητας του δεύτερου κόμβου.

- Υπολογισμός πιθανοτήτων

Χρησιμοποιώντας το Genie, μπορούν να υπολογιστούν πιθανότητες δίνοντας δεδομένα για κάποιους από τους γονεϊκούς κόμβους. Έτσι, διαλέγοντας μία κατάσταση (state) στους γονεϊκούς κόμβους και πατώντας το update [] υπολογίζεται η κατανομή της πιθανότητας με βάσει την επιλεγμένη κατάσταση του γονεϊκού κόμβου. Επίσης, κάνοντας δεξί κλικ σε κάποιον κόμβο, στην καρτέλα Value, φαίνεται γραφικά η πιθανότητα κάθε γεγονότος με βάση την επιλογή που έχει γίνει σε αντίστοιχο γονεϊκό κόμβο:

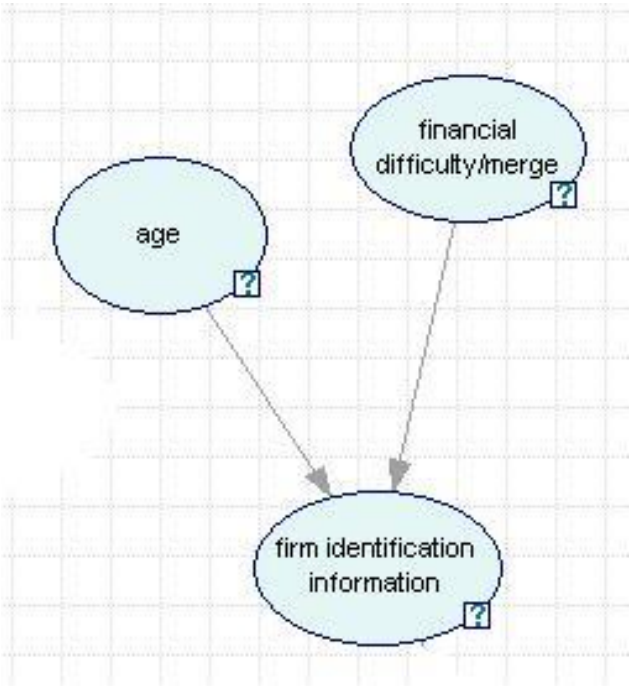


3.4 Σχεδιασμός του δικτύου Bayes

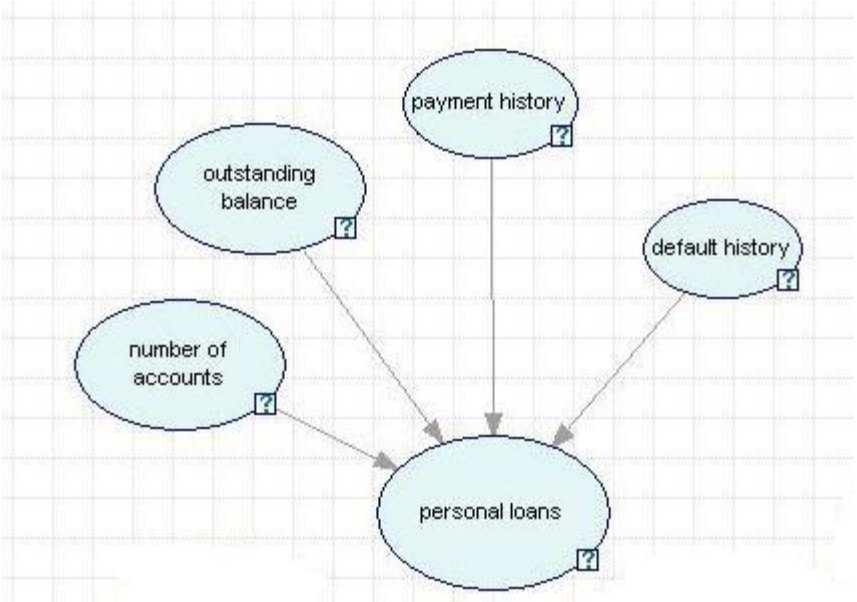
Με βάση το μοντέλο που επιλέχθηκε σύμφωνα με την έρευνα που περιγράφηκε στην παράγραφο 3.2, θα αναπαρασταθεί το μοντέλο με ένα δίκτυο Bayes.

Το δίκτυο αποτελείται από 17 κόμβους.

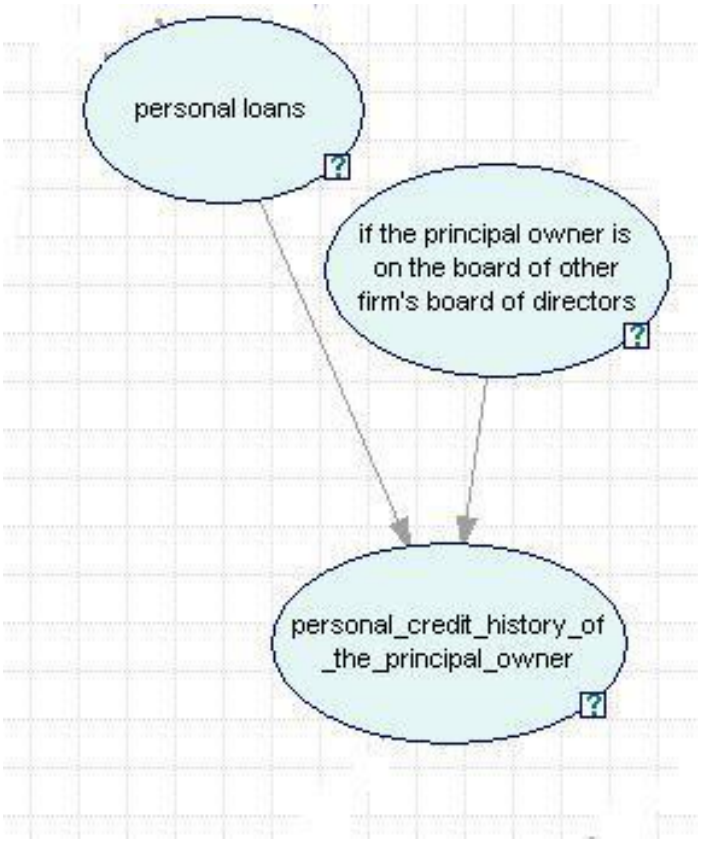
- Ο κόμβος “firm identification information” (πληροφορίες για την ταυτότητα μιας επιχείρησης) έχει γονείς τους κόμβους “age” (ηλικία) και “financial difficulty/merge” (η αποδοτικότητα της επιχείρησης ,καθώς και η οικονομική δυσκολία που ίσως αντιμετωπίζει).



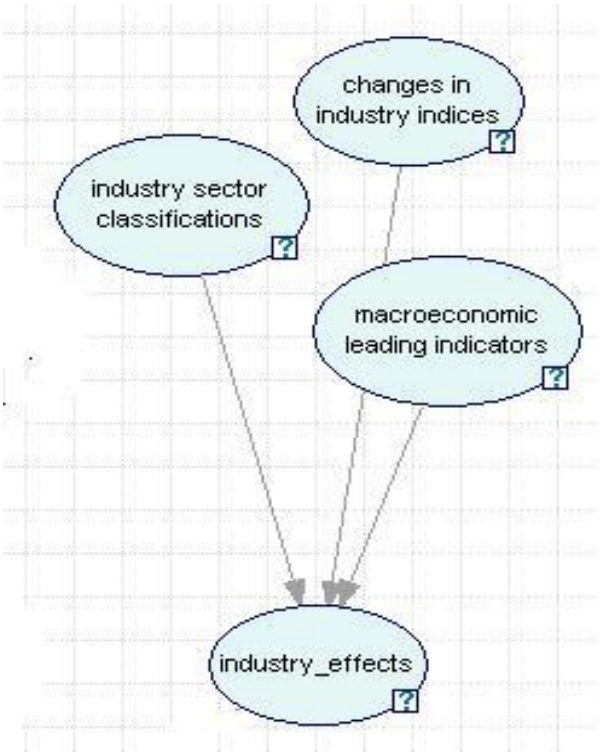
- Ο κόμβος “personal loans” (προσωπικά δάνεια) έχει γονείς τους κόμβους “number of accounts” (αριθμός λογαριασμών) , “outstanding balance” (το υπόλοιπο ενός δανείου), “payment history” (το ιστορικό πληρωμών) και “default history” (το ιστορικό αθέτησης αποπληρωμής ενός δανείου).



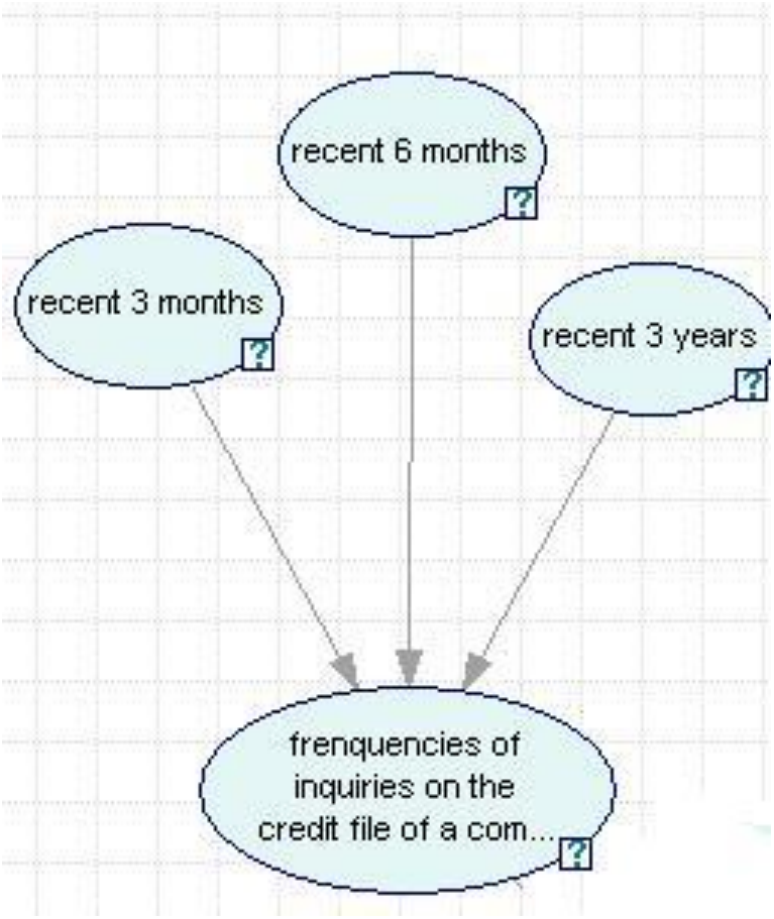
- Οι κόμβοι “personal loans” (προσωπικά δάνεια) και “if the principal owner is on the board of other firm's board of directors” (αν ο κύριος ιδιοκτήτης είναι στο συμβούλιο και άλλων διευθυντικών συμβουλίων άλλων επιχειρήσεων) είναι οι γονείς του κόμβου “personal credit history of the principal owner” (προσωπικό πιστωτικό ιστορικό του κύριου ιδιοκτήτη).



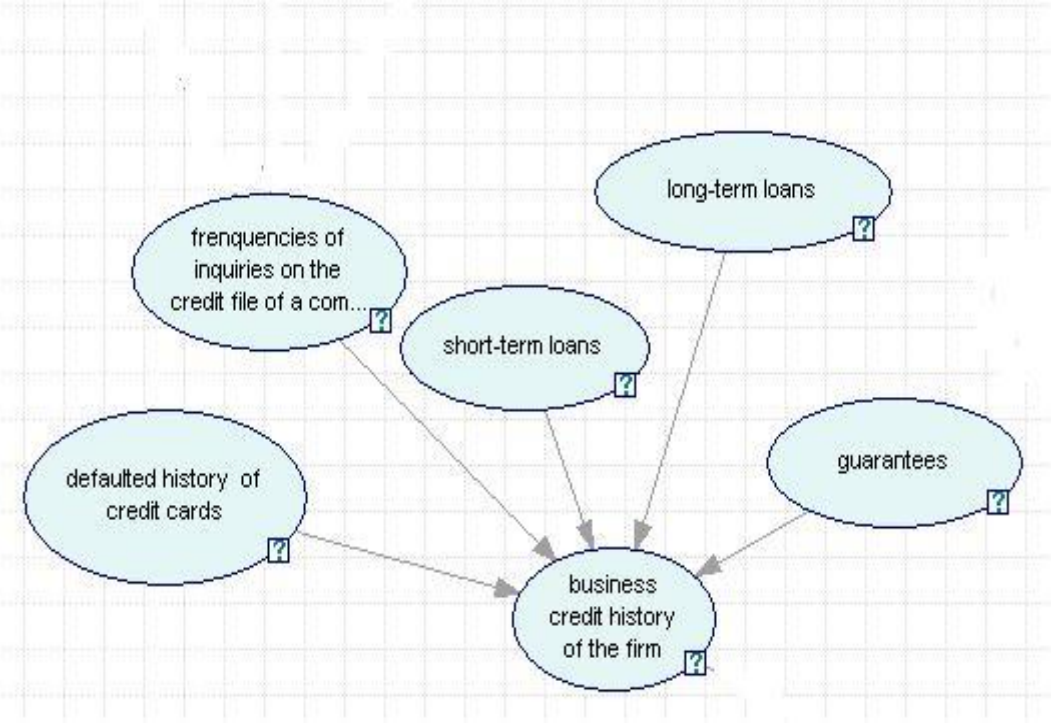
- Ο κόμβος “industry effects” (βιομηχανικές επιρροές) έχει γονείς τους κόμβους “industry sector classifications” (κατηγοριοποίηση του βιομηχανικού τομέα) , “changes in industry indices” (αλλαγές στους κλαδικούς δείκτες) και “macroeconomic leading indicators” (βασικοί μακροοικονομικοί δείκτες).



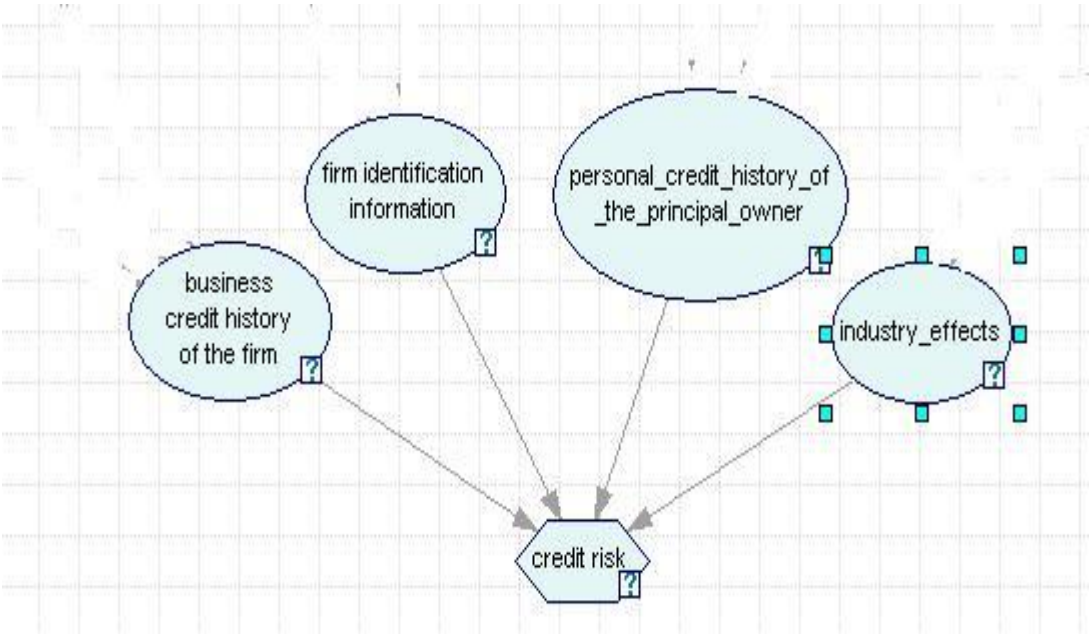
- Ο κόμβος “frenquencies of inquiries on the credit file of a company” (η συχνότητα των καταχωρήσεων στο πιστωτικό αρχείο) έχει γονείς τους κόμβους “recent 3 months” (τους τελευταίους 3 μήνες), “recent 6 months” (τους τελευταίους 6 μήνες) και “recent 3 years” (τα τελευταία 3 χρόνια).



- Ο κόμβος “business credit history of the firm” (το πιστωτικό ιστορικό μιας επιχείρησης) έχει γονείς τους κόμβους “defaulted history of credit cards” (ιστορικό αθέτησης αποπληρωμής πιστωτικών καρτών), “frenquencies of inquiries on the credit file of a company” (η συχνότητα των καταχωρήσεων στο πιστωτικό αρχείο), “short-term loans” (βραχυπρόθεσμα δάνεια), “long-term loans” (μακροπρόθεσμα δάνεια) και “guarantees” (εγγυήσεις).



- Τέλος ο κόμβος “credit risk”(πιστωτικό ρίσκο) αποτελεί το παιδί των κόμβων “firm identification information” (πληροφορίες για την ταυτότητα μιας επιχείρησης), “personal credit history of the principal owner” (προσωπικό πιστωτικό ιστορικό του κύριου ιδιοκτήτη), “industry effects” (βιομηχανικές επιρροές) και του “business credit history of the firm” (το πιστωτικό ιστορικό μιας επιχείρησης)



Το τελικό δίκτυο παρουσιάζεται παρακάτω.

credit risk
business credit history of the firm
short-term loans
frenquencies of inquiries on the credit file of a com...
defaulted history of credit cards
guarantees
long-term loans
firm identification information
age
financial difficulty/merge
personal credit history of the principal owner
personal loans
if the principal owner is on the board of other firm's board of directors
industry effects
industry sector classifications
changes in industry indices
macroeconomic leading indicators

Κεφάλαιο 4: Αποτελέσματα και μελλοντικές βελτιώσεις

4.1 Συμπεράσματα

Συνοψίζοντας μπορούμε να τονίσουμε τα θετικά των δικτύων Bayes και τους λόγους που είναι πιο αποτελεσματικά από τις υπόλοιπες μεθόδους στην αξιολόγηση της πιστοληπτικής ικανότητας κάποιου.

Ένα πολύ χαρακτηριστικό πλεονέκτημα των δικτύων Bayes είναι ότι ακόμη και αν δεν γνωρίζουμε όλα τα στοιχεία που παρουσιάζονται στο δίκτυο του μοντέλου μας ,μπορούμε να βγάλουμε συμπεράσματα για την πιστοληπτική ικανότητα κάποιου. Για παράδειγμα, το πιστωτικό ιστορικό μιας επιχείρησης επηρεάζεται από τους παράγοντες: ιστορικό αθέτησης αποπληρωμής πιστωτικών καρτών, συχνότητα των καταχωρήσεων στο πιστωτικό αρχείο, βραχυπρόθεσμα δάνεια, μακροπρόθεσμα δάνεια και εγγυήσεις. Εάν ένα χρηματοπιστωτικό ίδρυμα δεν γνωρίζει έστω τα βραχυπρόθεσμα δάνεια ενός υποψήφιου πελάτη του ,χρησιμοποιώντας το δίκτυο Bayes της προηγούμενης ενότητας μπορεί να αξιολογήσει κατά πόσο αυτό επηρεάζει το πιστωτικό ιστορικό μιας επιχείρησης και κατ'επέκταση το πιστωτικό ρίσκο του πελάτη. Επίσης τα δίκτυα Bayes κάνουν εμφανείς τις σχέσεις των μεταβλητών μεταξύ τους. Παρατηρώντας το δίκτυο μας είναι ξεκάθαρο ποιοι παράγοντες επηρεάζουν άμεσα το πιστωτικό ρίσκο του πελάτη. Αυτοί είναι οι πληροφορίες που έχουμε για την ταυτότητα της επιχείρησης, το προσωπικό πιστωτικό ιστορικό του κύριου ιδιοκτήτη, οι βιομηχανικές επιρροές και το πιστωτικό ιστορικό της επιχείρησης. Όμοια μπορούμε να δούμε ποιοι παράγοντες επηρεάζουν με τη σειρά τους παραπάνω παράγοντες. Αυτή η ξεκάθαρη εικόνα που έχουμε για τις σχέσεις των μεταβλητών μεταξύ τους ,μας διευκολύνει να κατανοήσουμε το πρόβλημα. Επίσης γνωρίζοντας τις αιτιώδεις σχέσεις μεταξύ τους μπορούμε να κάνουμε προβλέψεις κατά τη διάρκεια παρεμβάσεων. Για παράδειγμα, γνωρίζουμε ότι οι πληροφορίες για την ταυτότητα μιας επιχείρησης επηρεάζονται από την ηλικία της επιχείρησης

και την αποδοτικότητα της επιχείρησης ,καθώς και την οικονομική δυσκολία που ίσως αντιμετωπίζει. Αν γνωρίζουμε την ηλικία μιας επιχείρησης, τότε μπορούμε να κάνουμε προβλέψεις για το πόσο επηρεάζει αυτό το πιστωτικό ρίσκο του πελατή. Αν υποθέσουμε ότ είναι μια νέα επιχείρηση , τότε το πιστωτικό ρίσκο της είναι πιθανό να είναι μεγάλο καθώς υπάρχει αβεβαιότητα για την επιτυχία της επιχείρησης Επίσης τα δίκτυα Bayes διευκολύνουν το συνδυασμό γνωστικής περιοχής και δεδομένων. Σε κάποιες περιπτώσεις η γνώση των ειδικών είναι χρησιμότερη από τα συμπεράσματα κάποιων στατιστικών μεθόδων. Αυτό είναι κάτι που τα δίκτυα Bayes μπορούν να ενσωματώσουν.

4.2 Μελλοντικές Βελτιώσεις

Για περαιτέρω διερεύνηση της αξιοπιστίας και της ικανότητας αξιολόγησης της πιστοληπτικής ικανότητας κάποιου με τη βοήθεια του μοντέλου Bayes που παρουσιάστηκε στην προηγούμενη ενότητα, προτείνονται τα παρακάτω. Θα μπορούσαν να βρεθούν πραγματικά δεδομένα, έτσι ώστε να βρεθούν οι πιθανότητες που ενώνουν όλους τους κόμβους μεταξύ τους. Επίσης κάποιες από αυτές τις πιθανότητες θα μπορούσαν να οριστούν και σύμφωνα με τη γνώμη ειδικών οικονομολόγων που ασχολούνται με τους παράγοντες που επηρεάζουν την πιστοληπτική ικανότητα κάποιου. Έχοντας πλέον όλες τις πιθανοτικές σχέσεις που ενώνουν όλους τους κόμβους μεταξύ τους, το δίκτυο να δοκιμαστεί και να ελεγθεί αν οι μεταβλητές του μοντέλου εξυπηρετούν το σκοπό τους ή αν πρέπει να αφαιρεθούν. Επίσης αξιοποιώντας τη γνώση και εμπειρία των ειδικών, να εξεταστεί μήπως υπάρχουν κάποιες ακόμη μεταβλητές που πρέπει να προστεθούν στο μοντέλο.

Βιβλιογραφία

1. . http://www.euretirio.com/2011/06/pistoliptiki-ikanotita.html
2. Ben-Gal I., *Bayesian Networks*, in Ruggeri F., Faltin F. & Kenett R., Encyclopedia of Statistics in Quality & Reliability, Wiley & Sons (2007)
3. Todd A. Stephenson, *An introduction to Bayesian network theory and usage*, IDIAP-RR 00-03
4. Judea Pearl, Stuart Russell , *Bayesian Networks*, Department of Statistics Papers, Department of Statistics , UCLA, UCLA
5. David Heckerman, *A tutorial On Learning With Bayesian Networks*, Microsoft Research
6. Βασιλούδης θεόδωρος, *Ανάπτυξη Αλγορίθμων Ταξινόμησης Δεδομένων Πολλαπλών Ετικετών με χρήση Δικτύων Bayes,* Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Σχολή Θετικών Επιστημών, Τμήμα Πληροφορικής
7. Ray Tsaih, Yu-Jane Liu, Wenching Liu, Yu-Ling Lien, *Credit scoring system for small business loans*, Journal Decision Support Systems
8. http://genie.sis.pitt.edu/wiki/GeNIe_Documentation
9. Amos Tversky and Daniel Kahneman, *Judgment under Uncertainty: Heuristics and Biases, Science*
New Series, Vol. 185, No. 4157 (Sep. 27, 1974), pp. 1124-1131
10. Heckman et al.(1995)
Heckman, T., Dahlem, M., Lehnert, M., Fabbiano, G., Gilmore, D., & Waller, W. 1995, ApJ, 448, 98
11. Graham Cooper, *Cognitive load theory as an aid for instructional design, Australian Journal of Educational Technology 1990, 6(2), 108-113*
12. Murphy, K. (1998). *A brief introduction to graphical models and Bayesian networks*. http://www.cs. ubc.ca/~murphyk/Bayes/bnintro.html.
Earlier version appears at Murphy K. (2001) The Bayes Net Toolbox for Matlab, Computing Science and Statistics, 33, 2001.