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

Πρόβλεψη μελλοντικών σχέσεων μεταξύ των χρηστών των μέσων κοινωνικής δικτύωσης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΕΜΜΑΝΟΥΗΛ Ε. ΚΩΝΣΤΑΝΤΙΝΙΔΗΣ-ΓΕΩΡΓΙΟΥ

Επιβλέπων : Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2016

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

Πρόβλεψη μελλοντικών σχέσεων μεταξύ των χρηστών των μέσων κοινωνικής δικτύωσης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΕΜΜΑΝΟΥΗΛ Ε. ΚΩΝΣΤΑΝΤΙΝΙΔΗΣ-ΓΕΩΡΓΙΟΥ

Επιβλέπων : Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 13η Οκτωβρίου 2016.

Ανδρέας-Γεώργιος Σταφυλοπάτης Καθηγητής Ε.Μ.Π.	Γεώργιος Στάμου Επίκουρος Καθηγητής Ε.Μ.Π.	Γεώργιος Σιόλας Ε.ΔΙ.Π. Ε.Μ.Π.

Αθήνα, Οκτώβριος 2016

Εμμανουήλ Ε. Κωνσταντινίδης-Γεωργίου

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

Περίληψη

Ο σκοπός αυτής της διπλωματικής εργασίας είναι η μελέτη του προβλήματος της πρόβλεψης ακμών στα μέσα κοινωνικής δικτύωσης. Αρχικά, παρουσιάζονται τα βασικά συστατικά ενός κοινωνικού δικτύου και δίνεται ο ορισμός του προβλήματος. Επισημαίνεται η σημαντικότητα του προβλήματος σε ζητήματα της καθημερινότητας και τονίζονται τα προβλήματα που συναντούνται κατά την διαδικασία επίλυσης του.

Κατα την διαδικασία επίλυσης μελετούνται γραφοθεωρητικές τεχνικές, αλλά και πιο ευφυείς και περίπλοκες τεχνικές. Οι γραφοθεωρητικές τεχνικές βασίζονται στην μελέτη της δομής του δικτύου και την ομοιότητα μεταξύ των κόμβων. Δίνεται ιδιαίτερη βαρύτητα σε τεχνικές που μελετούν τους γειτόνους, τα μονοπάτια και τους τυχαίους περιπάτους μεταξύ δυο κόμβων του δικτύου.

Η επίλυση του προβλήματος με ευφυείς τεχνικές, απαιτεί τον ορισμό του αντίστοιχου προβλήματος δυαδικής ταξινόμησης, τον πειραματισμό με τις τιμές των παραμέτρων των διαφόρων ταξινομητών, τον σχεδιασμό και τη δημιουργία των δειγμάτων εκπαίδευσης και την αντιμετώπιση σημαντικών ζητημάτων που προκύπτουν. Στην παρούσα εργασία εξετάζεται η απόδοση πολλών διαφορετικών ταξινομητών.

Επίσης, γίνεται λεπτομερής αναφορά στα πειραματικά πρωτόκολλα που χρησιμοποιήθηκαν για τις γραφοθεωρητικές και τις ευφυείς τεχνικές, όπως και στον τρόπο με τον οποίο υπολογίζεται η απόδοση των διαφόρων τεχνικών πρόβλεψης. Τονίζεται ο λόγος χρήσης δυο διαφορετικών πειραματικών πρωτοκόλλων και γίνεται μια παρουσίαση της μεθόδου διασταυρούμενης αντεπικύρωσης, η οποία χρησιμοποιείται για τον υπολογισμό της απόδοσης όλων των τεχνικών.

Τέλος, παρουσιάζονται τα αποτελέσματα της έρευνας μας με την μορφή γραφικών παραστάσεων και σε μορφή τέτοια ώστε να μπορεί να γίνει γρήγορη και εύκολη σύγκριση μεταξύ των τεχνικών πρόβλεψης. Ακολουθεί η αξιολόγηση των τεχνικών και ο σχολιασμός των αποτελεσμάτων για τα σύνολα δεδομένων που χρησιμοποιήθηκαν. Η εργασία ολοκληρώνεται με την αναφορά στα συμπεράσματα που εξάγονται από την μελέτη του προβλήματος και τις μελλοντικές κατευθύνσεις.

Λέξεις κλειδιά

Μέσα κοινωνικής δικτύωσης, πρόβλεψη ακμών, γραφοθεωρητικές τεχνικές, ευφυείς τεχνικές, απόδοση, ταξινομητές, τυχαίοι περίπατοι, γράφος

Abstract

The aim of this thesis is to study the link prediction problem in social networks. Initially, we introduce the basic components of a social network and we define the problem. The importance of this problem is highlighted and we also point out the difficulties that we encounter during the solving process.

We study many graphtheoretical techniques but also more complex, machine learning techniques for solving this problem. Graphtheoretical techniques are based on examining the structure of the network and the similarity between nodes. We focus on techniques that examine the neighborhood, the paths and the random walks between two nodes of the network.

In order to solve the problem by using machine learning techniques, we first need to define the counterpart problem of binary classification, to experiment with the values of the classification parameters, to design and create the training examples for the classifiers and to deal with difficult problems that emerge during the classification process.

We also give a detailed report of the experimental setup that we use for solving the problem, but also a detailed report on how we measure the performance of different techniques. We highlight the importance of using two different experimental setups for the techniques that we study and we also introduce the idea of cross validation, a method that we use to measure the performance.

Finally, we present our results using charts, in a way that it is easier for the reader to find and compare the performance of different techniques. We evaluate our findings and comment on the results before we give our conclusions for our study and suggest future directions for the link prediction problem in social networks.

Key words

Online social networks, link prediction,graph-theoretical techniques, machine learning techniques, performance, classifier, random walks, graph

Ευχαριστίες

Με την παρούσα διπλωματική εργασία ολοκληρώνεται η ακαδημαϊκή μου πορεία στο Εθνικό Μετσόβιο Πολυτεχνείο και ταυτόχρονα μια σημαντική περίοδος της ζωής μου κλείνει τον κύκλο της. Για το λόγο αυτό θα ήθελα να ευχαριστήσω εγκαρδίως τα άτομα που με βοήθησαν να φτάσω μέχρι αυτό το σημείο.

Καταρχήν, οφείλω ένα μεγάλο ευχαριστώ στον κ. Ανδρέα-Γεώργιο Σταφυλοπάτη, Καθηγητή Ε.Μ.Π., για την ευκαιρία που μου προσέφερε να εκπονήσω αυτή τη διπλωματική εργασία, τον κ. Γεώργιο Αλεξανδρίδη, διδάκτορα Ε.Μ.Π. για την βοήθειά και την υπομονή του σε όλα τα στάδια της εργασίας και τους κ.κ. Γεώργιο Στάμου, Αναπληρωτή Καθηγητή ΕΜΠ, και Γεώργιο Σιόλα, Ε.ΔΙ.Π ΕΜΠ για την τιμή που μου κάνανε να είναι μέλη της επιτροπής εξέτασης της διπλωματικής μου εργασίας. Επίσης, θα ήθελα να ευχαριστήσω την οικογένειά μου, που μου συμπαραστάθηκε όλα αυτά τα χρόνια, καθώς και τον Παρασκευά.

Εμμανουήλ Ε. Κωνσταντινίδης-Γεωργίου,

Αθήνα, 13η Οκτωβρίου 2016

Η εργασία αυτή είναι επίσης διαθέσιμη ως Τεχνική Αναφορά , Εθνικό Μετσόβιο Πολυτεχνείο, Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών, Εργαστήριο Τεχνολογίας Λογισμικού, Οκτώβριος 2016.

URL: `http://www.softlab.ntua.gr/techrep/`
FTP: `ftp://ftp.softlab.ntua.gr/pub/techrep/`

Περιεχόμενα

Κατάλογος πινάκων

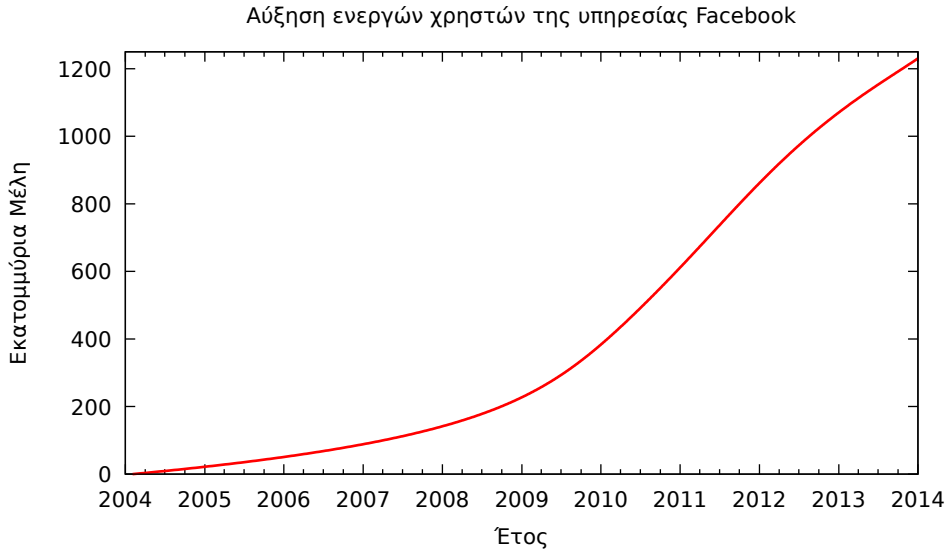
Κατάλογος σχημάτων

Κατάλογος Αλγορίθμων

Κεφάλαιο 1

Εισαγωγή

Τα *μέσα κοινωνικής δικτύωσης* (online social networks) είναι πλέον πολύ διαδεδομένα στην εποχή μας. Υπηρεσίες όπως το Facebook και το Twitter είναι γνωστές και χρησιμοποιούνται από ένα πολύ μεγάλο ποσοστό ανθρώπων (Σχήμα 1.1). Η απήχηση αυτών των υπηρεσιών οφείλεται στην ευκολία επικοινωνίας και ανταλλαγής πληροφορίας μεταξύ των χρηστών και υποστηρίζεται από την ραγδαία ανάπτυξη του Internet. Είναι σαφές, ότι η μεγάλη εξάπλωση των μέσων κοινωνικής δικτύωσης δεν θα μπορούσε να περάσει απαρατήρητη και από την επιστημονική κοινότητα.



Σχήμα 1.1: Η ραγδαία εξέλιξη των μέσων κοινωνικής δικτύωσης (Πηγή: http://en.wikipedia.org/wiki/Facebook)

Γενικότερα όμως, τις τελευταίες δεκαετίες τα *κοινωνικά δίκτυα* (social networks) αποτελούν αντικείμενο μελέτης για επιστήμονες από διάφορα επιστημονικά πεδία. Κοινωνιολόγοι, ψυχίατροι, φυσικοί, οικονομολόγοι και επιστήμονες των υπολογιστών μελετούν και αναλύουν τα κοινωνικά δίκτυα και πώς αυτά εξελίσσονται και αλλάζουν με την πάροδο του χρόνου. Ένας διαδεδομένος ορισμός των κοινωνικών δικτύων δίνεται ακριβώς παρακάτω [Wass94]

Ορισμός 1: *Ένα κοινωνικό δίκτυο αποτελείται από ένα σύνολο πεπερασμένου αριθμού δραστών καθώς και των σχέσεων που τους διέπουν. Επιπρόσθετα, πληροφορίες σχετικές με το είδος των δεσμών μεταξύ των δραστών αποτελούν απαραίτητο χαρακτηριστικό των κοινωνικών δικτύων.*

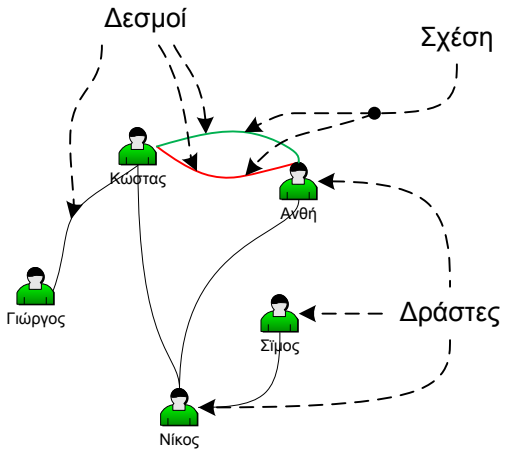
Ο συγκεκριμένος ορισμός εισάγει τρεις επιπρόσθετες έννοιες και πιο συγκεκριμένα αυτές του *δράστη* (actor), της *σχέσης* (relation) και του *δεσμού* (tie), οι ορισμοί των οποίων παρουσιάζονται στη συνέχεια [Wass94]

Ορισμός 2: *Ο όρος δράστης αναφέρεται σε μια κοινωνική οντότητα. Δράστες μπορεί να είναι μεμονωμένες, εταιρικές ή συλλογικές κοινωνικές μονάδες. Παραδείγματα δραστών είναι οι άνθρωποι σε*

ένα σύνολο, τα τμήματα μιας επιχείρησης, ένας δημόσιος οργανισμός σε μια πόλη ή οι χώρες του κόσμου. Ο όρος δράστης δεν υπονοεί απαραίτητα την δυνατότητα του δράστη να δρα, ενώ τα περισσότερα κοινωνικά δίκτυα αναφέρονται σε ένα σύνολο δραστών ιδίου τύπου.

Ορισμός 3: *Οι δράστες συνδέονται μεταξύ τους με δεσμούς. Το βασικό χαρακτηριστικό ενός δεσμού είναι ότι εγκαθιστά μια σύνδεση ανάμεσα σε ένα ζευγάρι δραστών. Παραδείγματα δεσμών αποτελούν η φιλία μεταξύ δυο ανθρώπων, η μεταφορά υλικών αγαθών μεταξύ δυο οργανισμών, η μετανάστευση από μια χώρα του κόσμου σε μια άλλη κ.ο.κ.*

Ορισμός 4: *Ο όρος σχέση αναφέρεται σε ένα σύνολο δεσμών ιδίου τύπου μεταξύ των μελών ενός συνόλου. Για παράδειγμα, σχέση μπορεί να αποτελούν όλες οι φιλίες μεταξύ ζευγαριών παιδιών σε μια σχολική τάξη ή το σύνολο διπλωματικών δεσμών μεταξύ των χωρών του κόσμου.*



Σχήμα 1.2: Στιγμιότυπο κοινωνικού δικτύου

Το Σχήμα 1.2 παρουσιάζει ένα στιγμιότυπο ενός κοινωνικού δικτύου που θα μας βοηθήσει να κατανοήσουμε τους ορισμούς των δραστών, των δεσμών και των σχέσεων. Σε αυτό το κοινωνικό δίκτυο το σύνολο των δραστών αποτελείται από πέντε ανθρώπους που συνδέονται μεταξύ τους με δεσμούς επικοινωνίας. Η γραμμή μεταξύ του Σίμου και του Νίκου αποτελεί ένα παράδειγμα δεσμού, ενώ παρατηρούμε ότι ο Κώστας και η Ανθή συνδέονται με δυο δεσμούς. Οι δυο αυτοί ξεχωριστοί δεσμοί επικοινωνίας αποτελούν μια σχέση.

Το αυξανόμενο ενδιαφέρον για τα κοινωνικά δίκτυα είχε ως αποτέλεσμα την ανάδυση ενός νέου ερευνητικού πεδίου, αυτού της *ανάλυσης κοινωνικών δικτύων* (social network analysis) [Wass94]. Οι ρίζες του πεδίου αυτού βρίσκονται στην κοινωνιολογία, την ανθρωπολογία, την ψυχολογία, τα μαθηματικά και την στατιστική. Βασικό χαρακτηριστικό του συγκεκριμένου πεδίου είναι η έμφαση που δίνεται στην μελέτη των σχέσεων μεταξύ των οντοτήτων του δικτύου, αντί της επεξεργασίας των χαρακτηριστικών των ίδιων των δραστών.

Το πεδίο της ανάλυσης κοινωνικών δικτύων αποτελεί τη βάση πάνω στην οποία ορίζονται και θεμελιώνονται θεωρητικές έννοιες και αξιολογούνται μοντέλα και θεωρίες. Αποτελεί το λεξιλόγιο με το οποίο περιγράφονται τεχνικές και διαδικασίες επίλυσης προβλημάτων που αναφέρονται σε κοινωνικές μονάδες που σχετίζονται μεταξύ τους.

Σε αντίθεση με άλλες επιστήμες που ασχολούνται με τα κοινωνικά δίκτυα, όπως λ.χ. η κοινωνιολογία, το πεδίο της ανάλυσης των δικτύων παρέχει τα εργαλεία για την αναπαράσταση των κοινωνικών δικτύων, τη μελέτη της δομή τους και την εύρεση των σημαντικότερων κοινωνικών δραστών. Επίσης, δίνεται η δυνατότητα να αναγνωριστούν μοτίβα στις σχέσεις μεταξύ των δραστών, χωρίς να απαιτείται η μελέτη των χαρακτηριστικών και της συμπεριφοράς μεμονωμένων μελών του δικτύου.

1.1 Παραδείγματα κοινωνικών δικτύων

Η μελέτη των κοινωνικών δικτύων είχε απασχολήσει την επιστημονική κοινότητα πολύ πριν την ανάπτυξη των δικτύων υπολογιστών. Παρακάτω δίνονται ορισμένα παραδείγματα τέτοιων δικτύων, ενώ ταυτόχρονα καταδεικνύεται γιατί είναι σημαντική η ανάλυση τους. Σε αυτό το σημείο αξίζει να τονιστεί ξανά πως ο όρος δράστης στα κοινωνικά δίκτυα δεν αναφέρεται μόνο σε ανθρώπους, αλλά γενικότερα σε κοινωνικές μονάδες, είτε αυτές μπορεί να είναι ένας δημόσιος οργανισμός, ένα τμήμα μιας επιχείρησης ή μια χώρα του κόσμου.

Ας υποθέσουμε ότι ενδιαφερόμαστε για την επιχειρησιακή συμπεριφορά διαφόρων οργανισμών σε μια μεγάλη πόλη, όπως για παράδειγμα τον τύπο και το μέγεθος της χρηματοδότησης μη κερδοσκοπικών οργανισμών [Gala85]. Η μελέτη των σχέσεων μεταξύ των επιχειρήσεων, όπου αυτές μπορεί να έχουν την μορφή συνεργασίας, της ανταλλαγής προϊόντων και υπηρεσιών ή φιλίας μεταξύ των μελών τους, μπορεί να βοηθήσει στην κατανόηση των αποφάσεων των επιχειρήσεων για από κοινού χρηματοδότηση άλλων οργανισμών. Σε αυτό το παράδειγμα, οι επιχειρήσεις αποτελούν τους δράστες του κοινωνικού δικτύου, ενώ οι σχέσεις μεταξύ των δραστών περιγράφηκαν παραπάνω.

Ως ένα δεύτερο παράδειγμα, υποθέτουμε έναν ψυχολόγο που μελετά τον τρόπο με τον οποίο ένα σύνολο ανθρώπων καταλήγει σε μια απόφαση. Ένα τέτοιο σύνολο θα μπορούσε να αποτελείται από ενόρκους οι οποίοι προσπαθούν να φτάσουν στην τελική απόφαση για μια δικαστική υπόθεση [Hast83]. Με την βοήθεια της ανάλυσης κοινωνικών δικτύων μπορούμε να εξετάσουμε τις σχέσεις μεταξύ των ενόρκων και πως αυτοί επηρεάζονται, οργανώνονται, σκέφτονται και τελικά αποφασίζουν. Η μελέτη της επικοινωνίας μεταξύ των μελών του συνόλου θα μας δώσει μια συνολικότερη εικόνα της τελικής απόφασης. Οι δράστες του δικτύου σε αυτήν την περίπτωση είναι οι ένορκοι.

Ένα ακόμα παράδειγμα γενικότερου δικτύου θα μπορούσε να αποτελέσει η παγκόσμια οικονομία. Οι δράστες του δικτύου είναι οι χώρες του κόσμου και οι σχέσεις μεταξύ των δραστών αναφέρονται στις διάφορες οικονομικές συνεργασίες μεταξύ των χωρών. Οι σχέσεις των χωρών μπορεί να ποικίλλουν από την ανταλλαγή προϊόντων και αγαθών μέχρι και την οικονομική ή στρατιωτική υποστήριξη. Οι μέθοδοι της ανάλυσης κοινωνικών δικτύων μπορεί να βοηθήσουν να αναγνωριστεί ποιες είναι οι ισχυρότερες οικονομικά χώρες, ποιες χώρες έχουν τις περισσότερες διπλωματικές σχέσεις και πως επηρεάζεται η παγκόσμια οικονομία από μια πιθανή διάλυση μιας σχέσης μεταξύ δυο χωρών.

Ως ένα τελευταίο παράδειγμα, θεωρούμε ένα *δίκτυο συγγραφής ερευνητικών εργασιών* (co-authorship network) για ένα συγκεκριμένο ερευνητικό πεδίο. Οι δράστες του δικτύου είναι οι συγγραφείς / ερευνητές, ενώ οι σχέσεις μεταξύ των συγγραφέων υποδηλώνουν συνεργασία και κοινή δημοσίευση μιας ερευνητικής εργασίας. Η ανάλυση ενός τέτοιου δικτύου μπορεί να μας βοηθήσει να αναγνωρίσουμε σημαντικούς ερευνητές του συγκεκριμένου ερευνητικού πεδίου με πολλές συνεργασίες και κοινές δημοσιεύσεις.

Το σημαντικότερο χαρακτηριστικό όλων των παραπάνω παραδειγμάτων είναι ότι δίνεται έμφαση στις σχέσεις μεταξύ των δραστών του κοινωνικού δικτύου και δεν μελετούνται οι δράστες και τα χαρακτηριστικά τους μεμονωμένα. Αυτή είναι η βασικότερη ιδέα του επιστημονικού πεδίου ανάλυσης των κοινωνικών δικτύων.

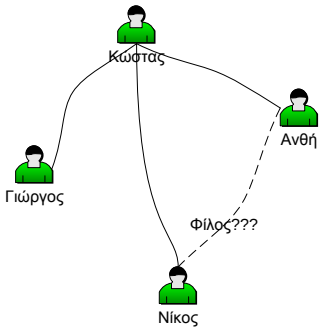
1.2 Το πρόβλημα της πρόβλεψης ακμών

Το ερευνητικό πεδίο της ανάλυσης κοινωνικών δικτύων περιλαμβάνει πολλά προβλήματα. Ένα από τα σημαντικότερα προβλήματα είναι αυτό της *πρόβλεψης ακμών* (link prediction), του οποίου ο ορισμός παρατίθεται στη συνέχεια [Libe03]

Ορισμός 5: *Δοσμένου ενός στιγμιότυπου ενός κοινωνικού δικτύου σε χρόνο t, προσπαθούμε να προβλέψουμε με ακρίβεια τις ακμές που θα προστεθούν στο δίκτυο στο χρονικό διάστημα από τη χρονική στιγμή t μέχρι τη στιγμή t'*

Ο ορισμός αυτός αναφέρεται στην εξέλιξη του δικτύου με την πάροδο του χρόνου, αλλά το πρόβλημα της πρόβλεψης ακμών θα μπορούσε να διατυπωθεί και για ένα στατικό στιγμιότυπο ενός δικτύου [Libe03]:

Ορισμός 6: *Δοσμένου ενός στατικού στιγμιότυπου ενός κοινωνικού δικτύου και χρησιμοποιώντας πληροφορίες σχετικά με την δομή του, προσπαθούμε να προβλέψουμε επιπλέον ακμές που δεν είναι ορατές στο δίκτυο ή που είναι πιθανόν να υπάρχουν*



Σχήμα 1.3: Το πρόβλημα της πρόβλεψης ακμών

Στο Σχήμα 1.3 παρουσιάζεται ένα απλό παράδειγμα που βοηθά στην κατανόηση του προβλήματος της πρόβλεψης ακμών. Στην εικόνα φαίνεται ένα στιγμιότυπο ενός περιορισμένου κοινωνικού δικτύου, όπου με συνεχόμενη γραμμή δηλώνεται η φιλία μεταξύ χρηστών. Για παράδειγμα, ο Κώστας είναι φίλος με την Ανθή, όπως και ο Γιώργος με την Ανθή. Το πρόβλημα της πρόβλεψης ακμών απαντάει στο ερώτημα του ποια είναι η πιθανότητα η Ανθή να είναι φίλη με τον Νίκο, μια ακμή που στο συγκεκριμένο στιγμιότυπο του δικτύου δεν φαίνεται να υπάρχει.

Το πρόβλημα της πρόβλεψης ακμών παρουσιάζει μεγάλο ενδιαφέρον, για πολλούς λόγους. Εξετάζοντας ξανά τα παραδείγματα της Ενότητας 1.1, επανερχόμαστε σε αυτό της χρηματοδότησης των μη κερδοσκοπικών οργανισμών από τις επιχειρήσεις μιας πόλης. Η επίλυση του προβλήματος της πρόβλεψης ακμών σε ένα τέτοιο δίκτυο θα μπορούσε να αναγνωρίσει νέες μελλοντικές χρηματικές συναλλαγές μεταξύ των επιχειρήσεων, μέσω της μελέτης της δομής του δικτύου και των σχέσεων των οργανισμών.

Στο παράδειγμα του δικτύου της παγκόσμιας οικονομίας, όπου οι χώρες του κόσμου αποτελούν τους δράστες του δικτύου και οι οικονομικές συναλλαγές αποτελούν τις σχέσεις μεταξύ των δραστών, το πρόβλημα της πρόβλεψης ακμών μπορεί να βοηθήσει στο να προταθούν νέες οικονομικές συνεργασίες για μια συγκεκριμένη χώρα του κόσμου. Επιπλέον, μπορεί να αποτιμηθεί η μελλοντική εικόνα της παγκόσμιας οικονομίας, μέσω της πρόβλεψης πιθανών οικονομικών συναλλαγών μεταξύ χωρών του κόσμου. Τέλος, στο δίκτυο συγγραφέων/ερευνητών ενός επιστημονικού πεδίου η επίλυση του προβλήματος της πρόβλεψης ακμών μπορεί να προτείνει μελλοντικές συνεργασίες και κοινές δημοσιεύσεις μεταξύ δυο ερευνητών.

Εκτός από τις παραπάνω σημαντικές χρήσεις, το πρόβλημα της πρόβλεψης ακμών βρίσκει εφαρμογή σε πολλά άλλα προβλήματα. Στην παρακάτω λίστα παρουσιάζονται συνοπτικά μια σειρά από σημαντικές εφαρμογές του εν λόγω προβλήματος

1. Η μελέτη πιθανών διασυνδέσεων τρομοκρατικών οργανώσεων μπορεί να γίνει με την βοήθεια της πρόβλεψης ακμών σε αντίστοιχα δίκτυα τρομοκρατικών οργανώσεων [Libe03].

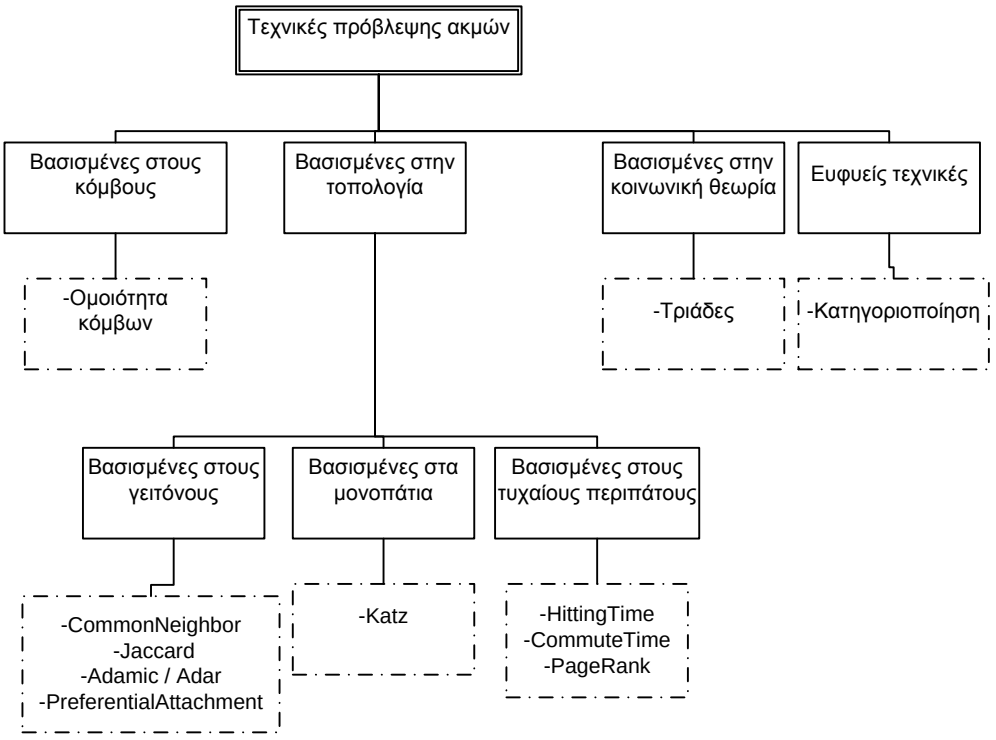
2. Η πρόταση νέων αντικειμένων στους χρήστες των *συστημάτων συστάσεων* (recommender systems) μπορεί να πραγματοποιηθεί και με χρήση τεχνικών πρόβλεψης ακμών [Li09, Huan05]. Σε αυτή την περίπτωση, ο γράφος του δικτύου αποτελείται από κόμβους που αντιστοιχούν σε χρήστες και αντικείμενα, ενώ οι ακμές δηλώνουν τη σχέση προτίμησης ενός χρήστη προς ένα αντικείμενο. Η πρόβλεψη μιας νέας ακμής ισοδυναμεί με την πρόταση ενός νέου αντικειμένου σε ένα χρήστη.

3. Τέλος, η πρόβλεψη ακμών μπορεί να χρησιμοποιηθεί στο πεδίο του Παγκοσμίου Ιστού, σε λειτουργίες όπως η αυτόματη δημιουργία υπερσυνδέσμων ιστοσελίδων [Adaf05], αλλά και στο

πεδίο της βιοϊατρικής, στην πρόβλεψη ακμών σε δίκτυα αλληλεπίδρασης πρωτεϊνών [Airo08] κ. ά.

1.3 Τεχνικές πρόβλεψης ακμών

Τις τελευταίες δεκαετίες έχουν πραγματοποιηθεί μελέτες με θέμα την πρόβλεψη ακμών στα κοινωνικά δίκτυα και έχουν προταθεί πολλές τεχνικές πρόβλεψης. Αρχικά χρησιμοποιήθηκαν γραφοθεωρητικές τεχνικές για να λύσουν το πρόβλημα, αναλύοντας κυρίως την δομή του κοινωνικού δικτύου [Libe03]. Αν και αυτές οι τεχνικές πλέον θεωρούνται πολύ απλές, αποτέλεσαν μια καλή αρχή για την μελέτη του συγκεκριμένου προβλήματος. Επίσης, εκτός από την συστηματική μελέτη των απλών, γραφοθεωρητικών τεχνικών, προτάθηκαν και σύγχρονες, ευφυείς τεχνικές πρόβλεψης που επιτυγχάνουν πολύ καλύτερη απόδοση [Wang14].



Σχήμα 1.4: Οι τεχνικές πρόβλεψης ακμών

Το Σχήμα 1.4 συνοψίζει τους τρόπους με τους οποίους μπορούν να ταξινομηθούν οι τεχνικές πρόβλεψης ακμών. Σε ένα πρώτο επίπεδο διακρίνονται οι μέθοδοι *βασισμένοι στους κόμβους* (node-based), οι μέθοδοι βασισμένες στην *τοπολογία του γράφου* (topology-based), οι μέθοδοι βασισμένες στην *κοινωνική θεωρία* (social-theory) και τέλος οι *ευφυείς μέθοδοι* (learning-based).

Από τα παραπάνω γίνεται εμφανές ότι το πρόβλημα της πρόβλεψης ακμών μπορεί να ιδωθεί είτε από τη σκοπιά της γραφοθεωρίας (λ.χ. μελετώντας ιδιότητες όπως ο βαθμός ενός κόμβου ή το μήκος των μονοπατιών μεταξύ δύο κόμβων) είτε σαν ένα *πρόβλημα δυαδικής κατηγοριοποίησης* (binary classification problem) από τη σκοπιά των ευφυών τεχνικών, όπου το ζητούμενο είναι το σύστημα να μπορεί να προβλέψει πότε ένα ζεύγος κόμβων συνδέεται με ακμή και πότε όχι. Η μελέτη αυτών των τεχνικών στο πλαίσιο των μέσων κοινωνικής δικτύωσης αποτελεί και το αντικείμενο της συγκεκριμένης διπλωματικής εργασίας, με την κάθε μια τεχνική που αναφέρεται εδώ να αναλύεται διεξοδικότερα στα επόμενα Κεφάλαια.

Τέλος, οι τεχνικές που βασίζονται στην τοπολογία του γράφου μπορούν να ταξινομηθούν περαιτέρω σε μεθόδους βασισμένες στους *γείτονες* (neighbor-based), σε μεθόδους βασισμένες στα *μονοπάτια* (path-based) και σε μεθόδους βασισμένες στους *τυχαίους περιπάτους* (random walks).

1.4 Κυριότερες προκλήσεις

Παρά την σπουδαιότητα που έχει η επίλυση του προβλήματος της πρόβλεψης ακμών στα μέσα κοινωνικής δικτύωσης, οι δυσκολίες που πρέπει να αντιμετωπιστούν είναι πολλές.

Τα μέσα κοινωνικής δικτύωσης είναι οντότητες που εξελίσσονται, αναπτύσσονται και καταστρέφονται με την πάροδο του χρόνου. Οι μεταβολές αυτές οφείλονται στην δυναμική συμπεριφορά των μελών των εν λόγω δικτύων. Για παράδειγμα, σε ένα μέσο κοινωνικής δικτύωσης όπως είναι το Facebook, νέοι χρήστες εγγράφονται στην υπηρεσία καθημερινά με αποτέλεσμα να δημιουργούνται νέοι κόμβοι του δικτύου διαρκώς. Επιπλέον, νέες φιλίες δημιουργούνται μεταξύ των χρηστών, που ισοδυναμούν με νέες ακμές του δικτύου. Το Σχήμα 1.1 δείχνει την αύξηση των ενεργών χρηστών της υπηρεσίας Facebook με την πάροδο του χρόνου και τονίζει με χαρακτηριστικό τρόπο πως τα μέσα κοινωνικής δικτύωσης είναι σε πολύ μεγάλο βαθμό δυναμικά. Το χαρακτηριστικό αυτό καθιστά την επίλυση του προβλήματος της πρόβλεψης ακμών εξαιρετικά δύσκολη.

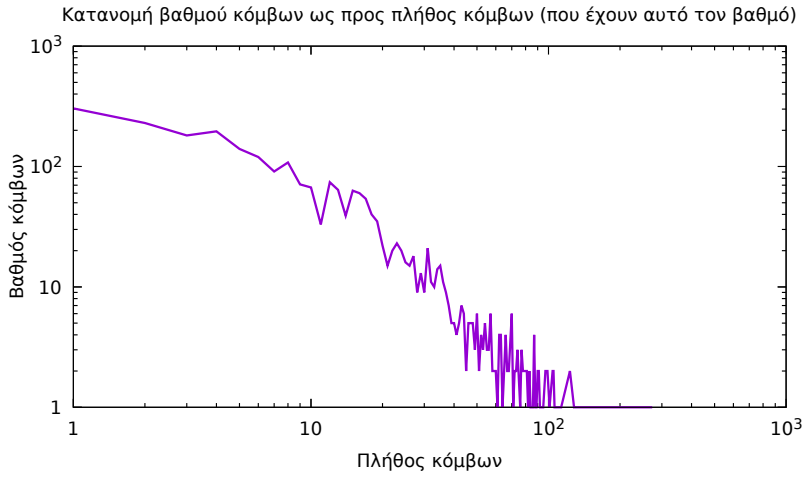
Ένα ακόμα πρόβλημα είναι το μέγεθος των μέσων κοινωνικής δικτύωσης. Το δίκτυο της υπηρεσίας του Facebook που προαναφέρθηκε, έχει εκατομμύρια κόμβους και δισεκατομμύρια ακμές μεταξύ των χρηστών. Γίνεται συνεπώς αντιληπτό ότι η περιγραφή και αποθήκευση μεγάλων κοινωνικών δικτύων του πραγματικού κόσμου είναι δύσκολη, ενώ η ανάλυση αυτών με τους σύγχρονους υπολογιστικούς πόρους είναι πρακτικά αδύνατη.

Επίσης, ένα ιδιαίτερο χαρακτηριστικό των μέσων κοινωνικής δικτύωσης είναι η *αραιότητα των ακμών τους* (edge sparsity). Ο αριθμός των υπαρκτών ακμών του δικτύου είναι σημαντικά μικρότερος από τον μέγιστο δυνατό αριθμό ακμών του αντίστοιχου *πλήρους γράφου* (clique) του ίδιου δικτύου, κρατώντας την πυκνότητα τους σε πολύ χαμηλά επίπεδα. Αν σκεφτούμε για παράδειγμα το δίκτυο της υπηρεσίας Facebook που προαναφέραμε, η πιθανότητα δυο τυχαία άτομα του δικτύου να γνωρίζονται μεταξύ τους είναι πάρα πολύ μικρή. Επιπρόσθετα, σε επίπεδο κόμβων, παρατηρείται το φαινόμενο ορισμένοι, ελάχιστοι, κόμβοι να έχουν πολύ μεγάλο αριθμό προσπιπτουσών ακμών, ενώ αντίθετα το συντριπτικά μεγαλύτερο ποσοστό των κόμβων του δικτύου έχει μικρό αριθμό προσπιπτουσών ακμών. Για παράδειγμα, στο Twitter, οι λογαριασμοί γνωστών καλλιτεχνών, αθλητών, πολιτικών κ. ά έχουν πολύ περισσότερους ακολούθους απ' ότι ο μέσος χρήστης της υπηρεσίας.

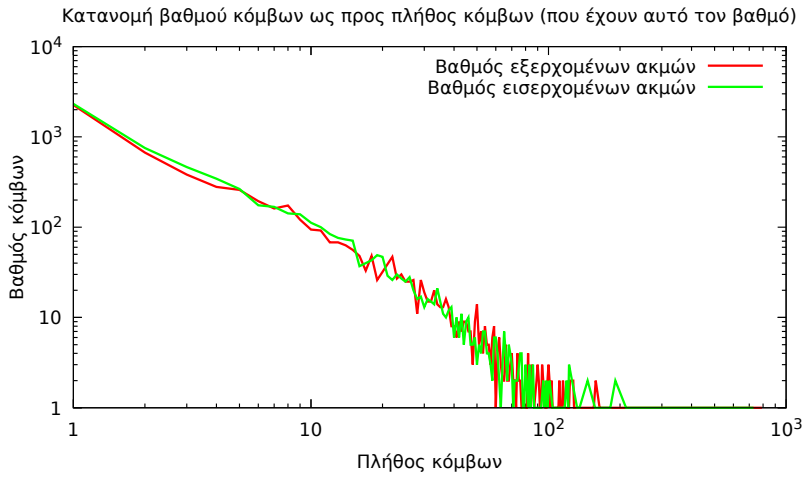
Το Σχήμα 1.5 απεικονίζει, σε λογαριθμικούς άξονες, την κατανομή του *βαθμού του κάθε κόμβου* (node degree) ως προς το πλήθος των κόμβων που έχουν τον συγκεκριμένο βαθμό για τις δύο συλλογές δεδομένων (datasets) που χρησιμοποιήθηκαν στο πειραματικό σκέλος της παρούσας εργασίας. Είναι χαρακτηριστική η γραμμική μείωση του βαθμού των κόμβων προς τον πλήθος των κόμβων που έχουν τον συγκεκριμένο βαθμό, στην λογαριθμική κλίμακα που παρατηρείται, μείωση που αποτελεί την χαρακτηριστική ιδιότητα των *δικτύων ελεύθερης κλίμακας* (scale-free networks).

Εξαιτίας του χαμηλού βαθμού που έχουν οι περισσότεροι κόμβοι του δικτύου, αναμένεται ότι η απόδοση των τεχνικών που βασίζονται στην γραφοθεωρία, θα έχουν σχετικά χαμηλή απόδοση. Επίσης, γίνεται αντιληπτό ότι ο συνεπαγόμενος *πίνακας γειτνίασης* (adjacency matrix) του εν λόγω δικτύου θα είναι πολύ αραιός, με λίγα μη-μηδενικά στοιχεία. Η αποθήκευση, η επεξεργασία και ο χειρισμός τέτοιων αραιών, μεγάλων πινάκων αποτελεί πρόκληση ακόμα και για τα σύγχρονα υπολογιστικά συστήματα.

Η αραιότητα των κοινωνικών δικτύων όμως, δεν προκαλεί προβλήματα μόνο στις γραφοθεωρητικές τεχνικές πρόβλεψης ακμών. Ένα πολύ σημαντικό πρόβλημα των ευφυών τεχνικών πρόβλεψης, είναι αυτό της μεροληπτικής μάθησης των ταξινομητών. Στα προβλήματα δυαδικής κατηγοριοποίησης, το σύνολο εκπαίδευσης των ταξινομητών είναι επιθυμητό να είναι κατά το δυνατό ισοκατανεμημένο, έτσι ώστε κατά την διαδικασία μάθησης το μοντέλο προβλέψεων να εκτεθεί και στις δυο κλάσεις του προβλήματος εξίσου. Συγκεκριμένα, στο πρόβλημα της πρόβλεψης ακμών στα κοινωνικά δίκτυα το σύνολο εκπαίδευσης θα αποτελείται από ζεύγη κόμβων που είτε υπάρχει ακμή μεταξύ τους και άρα ανήκουν στην θετική κλάση, είτε δεν υπάρχει ακμή να τους συνδέει και άρα ανήκουν στην αρνητική κλάση. Την έξοδο των ταξινομητών θα αποτελεί η κλάση των ζευγών κόμβων που βρίσκονται στο σύνολο ελέγχου. Λόγω της αραιότητας των κοινωνικών δικτύων δημιουργείται ένα παραμορφωμένο σύνολο εκπαίδευσης (class imbalance), όπου η κλάση που αναφέρεται στις υπαρκτές ακμές του δικτύου (θετική κλάση) υποεκπροσωπείται. Αντίθετα, τα δείγματα που ανήκουν στην κλάση που



(a) `soc-hamsterster`



(b) `soc-advogato`

Σχήμα 1.5: Γραμμική μείωση του βαθμού των κόμβων προς το πλήθος των κόμβων. Χαρακτηριστική γραφική παράσταση των δικτύων ελεύθερης κλίμακας

αναφέρεται στις μη-υπαρκτές ακμές του δικτύου (αρνητική κλάση) είναι υπερβολικά πολλά.

Η έμπρακτη εφαρμογή των ευφυών τεχνικών σε ένα πραγματικό πρόβλημα, όπως αυτό της πρόβλεψης ακμών, καθώς και αναζήτηση λύσεων στις δυσκολίες που συναντήσαμε, όπως το πρόβλημα της υποεκπροσώπησης της μιας κλάσης που περιγράφηκε παραπάνω, ήταν τα βασικότερα κίνητρα επιλογής για την εκπόνηση της συγκεκριμένης διπλωματικής εργασίας. Είναι επίσης σημαντικό να τονίσουμε, ότι ο λόγος που ασχοληθήκαμε με το πρόβλημα της πρόβλεψης μελλοντικών ακμών στα μέσα κοινωνικής δικτύωσης είναι η ύπαρξη μεγάλου αριθμού διαθέσιμων συλλογών δεδομένων (datasets) τέτοιων δικτύων με κατάλληλο μέγεθος ώστε να μπορούμε να εξάγουμε ασφαλή συμπεράσματα κατά την διαδικασία εκτέλεσης της πειραματικής διαδικασίας.

1.5 Δομή της εργασίας

Έχοντας ολοκληρώσει την εισαγωγή στο πρόβλημα που μελετούμε, σε αυτήν την ενότητα θα περιγράψουμε πώς δομείται η συγκεκριμένη διπλωματική εργασία.

Στο Κεφάλαιο 2, αναλύουμε τις γραφοθεωρητικές τεχνικές επίλυσης του προβλήματος της πρόβλεψης ακμών. Πιο συγκεκριμένα, αναφερόμαστε σε τεχνικές μέτρησης της ομοιότητας κόμβων του δικτύου μέσω της ανάλυσης των χαρακτηριστικών τους, σε τεχνικές που μελετούν την δομή του δι-

κτύου, στους τυχαίους περιπάτους και σε τεχνικές που στηρίζονται στην κοινωνική θεωρία.

Στο Κεφάλαιο 3 αναφερόμαστε στις ευφυείς τεχνικές επίλυσης, παρουσιάζοντας διάφορα είδη ταξινομητών. Μελετούμε τον τρόπο λειτουργίας κάθε ταξινομητή, απεικονίζουμε με απλά παραδείγματα την διαδικασία εκπαίδευσης και τονίζουμε τα πλεονεκτήματα και μειονεκτήματα χρήσης τους.

Στο Κεφάλαιο 4, αρχικά δίνουμε πληροφορίες για τα σύνολα δεδομένων που χρησιμοποιήσαμε. Στην συνέχεια, αναφερόμαστε στην μέθοδο διασταυρωμένης επικύρωσης και την χρησιμότητα της στο πρόβλημα μας. Περιγράφουμε λεπτομερώς την πειραματική διαδικασία που ακολουθήσαμε κατά την επίλυση του προβλήματος με χρήση των γραφοθεωρητικών και των ευφυών τεχνικών και αναλύουμε τις μετρικές που χρησιμοποιήσαμε για τον υπολογισμό της απόδοσης των τεχνικών.

Στο Κεφάλαιο 5, παρουσιάζουμε τα αποτελέσματα των πειραμάτων που πραγματοποιήσαμε, δίνοντας παράλληλα και τις παρατηρήσεις μας για την απόδοση των διαφόρων τεχνικών. Στην συνέχεια, επικεντρωνόμαστε στις ευφυείς τεχνικές, εξετάζοντας και σχολιάζοντας την απόδοση διαφόρων ταξινομητών για διαφορετικές τιμές των παραμέτρων τους.

Τέλος, στο Κεφάλαιο 6 παρουσιάζουμε γενικότερα συμπεράσματα που εξαγάγουμε από την μελέτη και επίλυση του προβλήματος της πρόβλεψης μελλοντικών ακμών στα μέσα κοινωνικής δικτύωσης και προτείνουμε μελλοντικές ιδέες και κατευθύνσεις για την προσέγγιση του προβλήματος.

Κεφάλαιο 2

Γραφοθεωρητικές τεχνικές επίλυσης του προβλήματος πρόβλεψης ακμών

Σε αυτό το κεφάλαιο θα μελετηθούν οι γραφοθεωρητικές τεχνικές που λύνουν το πρόβλημα της πρόβλεψης ακμών στα κοινωνικά δίκτυα. Οι τεχνικές αυτές πηγάζουν από πολλά διαφορετικά επιστημονικά πεδία, όπως τα μαθηματικά και οι κοινωνικές επιστήμες και στις περισσότερες περιπτώσεις δεν υπολογίζουν την ομοιότητα μεταξύ δυο κόμβων ενός γράφου και για αυτόν τον λόγο πρέπει να διαφοροποιηθούν. Χρονικά, αυτές ήταν οι πρώτες τεχνικές που προτάθηκαν ως τρόπος προσέγγισης του προβλήματος της πρόβλεψης ακμών [Libe03]. Βασικό χαρακτηριστικό όλων των τεχνικών (πλην των τυχαίων περιπάτων) που θα παρουσιαστούν παρακάτω είναι ο ντετερμινισμός.

Κάθε τεχνική υπολογίζει μια τιμή για κάθε μη υπαρκτή ακμή μεταξύ των κόμβων x και y του εξεταζόμενου γράφου. Υψηλή τιμή σημαίνει μεγάλη πιθανότητα μια ακμή να αποτελέσει μελλοντική ακμή του γράφου, ενώ αντίθετα χαμηλή τιμή, δηλώνει χαμηλή ομοιότητα μεταξύ των συγκεκριμένων κόμβων και άρα μια μη πιθανή ακμή στον γράφο.

2.1 Τεχνικές βασισμένες στους κόμβους

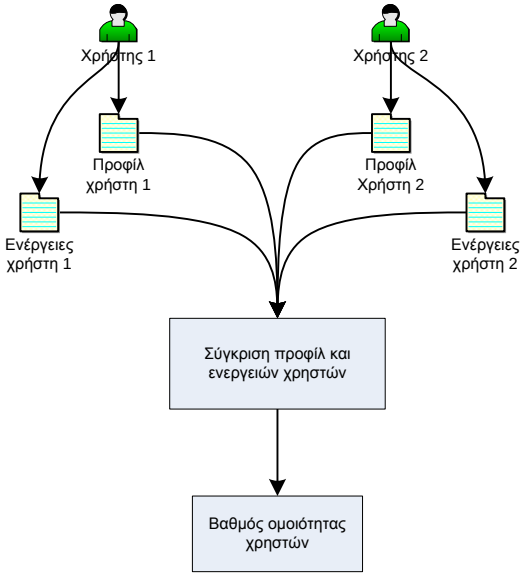
Οι τεχνικές που βασίζονται στους κόμβους, εξετάζουν πληροφορίες που είναι διαθέσιμες για κάθε χρήστη ενός μέσου κοινωνικής δικτύωσης και προσπαθούν να υπολογίσουν έναν βαθμό ομοιότητας μεταξύ χρηστών, χωρίς να εξετάζουν την τοπολογία του δικτύου. Αυτό γίνεται μελετώντας τα διάφορα χαρακτηριστικά και τις ενέργειες των χρηστών.

Κάθε χρήστης ενός μέσου κοινωνικής δικτύωσης έχει ένα προφίλ, το οποίο μπορεί να αναπαρασταθεί ως ένα διάνυσμα χαρακτηριστικών σε μια βάση δεδομένων. Το διάνυσμα αυτό μπορεί να περιέχει πληροφορίες όπως την ηλικία, το ηλεκτρονικό ταχυδρομείο, την περιοχή που διαμένει ο χρήστης, θρησκευτικές πεποιθήσεις, ενδιαφέροντα και χόμπι, αγαπημένες ταινίες, βιβλία και τραγούδια κ.λ.π. Το διάνυσμα αυτό είναι μοναδικό για κάθε χρήστη και τον χαρακτηρίζει σαφώς [Bhat11].

Επιπλέον, οι διάφορες ενέργειες του χρήστη μπορούν και αυτές να αποθηκευτούν σε μια βάση δεδομένων και να αποτελέσουν σημαντική πληροφορία για την πρόβλεψη ακμών. Για παράδειγμα, στο Facebook ενέργειες όπως το άκουσμα ενός τραγουδιού, το "Like" σε μια φωτογραφία και η συμμετοχή σε ένα μουσικό γεγονός, μπορούν να χρησιμοποιηθούν ως μέτρο σύγκρισης με τις προτιμήσεις άλλων χρηστών [Ande12].

Έτσι λοιπόν, μια απλή σύγκριση μεταξύ των διανυσμάτων χαρακτηριστικών δυο χρηστών μπορεί να μας δώσει μια τιμή, δηλαδή έναν βαθμό ομοιότητας μεταξύ των συγκεκριμένων χρηστών. Εφόσον, τα χαρακτηριστικά είναι συνήθως σε μορφή κειμένου, *τεχνικές βασισμένες σε κείμενο* (text-based techniques) μπορούν να χρησιμοποιηθούν για την εξαγωγή του επιθυμητού βαθμού ομοιότητας.

Για την καλύτερη κατανόηση των παραπάνω, το Σχήμα 2.1 παρουσιάζει την διαδικασία με την οποία μπορεί να βρεθεί ο βαθμός ομοιότητας μεταξύ δυο κόμβων χρηστών. Είναι σαφές, πως η ύπαρξη μεγάλου αριθμού ελλιπώς συμπληρωμένων προφίλ χρηστών θα έχει μεγάλη αρνητική επίδραση στην απόδοση των τεχνικών βασισμένων στους κόμβους.



Σχήμα 2.1: Διαδικασία υπολογισμού βαθμού ομοιότητας με τεχνικές βασισμένες στους κόμβους

2.2 Τεχνικές βασισμένες στην τοπολογία

Στις περιπτώσεις που το διάνυσμα χαρακτηριστικών των χρηστών δεν είναι διαθέσιμο, βρίσκουν εφαρμογή οι τεχνικές βασισμένες στην τοπολογία του δικτύου. Οι τεχνικές βασισμένες στην τοπολογία, όπως και πριν, υπολογίζουν μια τιμή για κάθε μη υπαρκτή ακμή του δικτύου. Το Σχήμα 2.2 απεικονίζει πώς λαμβάνοντας ως είσοδο το στιγμιότυπο ενός δικτύου και με χρήση τεχνικών που μελετούν την δομή του γράφου, μπορούν να προβλεφθούν μελλοντικές ακμές.

Στις παρακάτω τεχνικές με $\Gamma(x)$ συμβολίζεται το σύνολο που περιέχει όλους τους γείτονες του κόμβου x (αυτούς δηλαδή με τους οποίους συνδέεται άμεσα με ακμή), ενώ με $|\Gamma(x)|$ το πλήθος τους.

Σχήμα 2.2: Μελέτη τεχνικών βασισμένων στην τοπολογία του γράφου

2.2.1 Τεχνικές βασισμένες στους γειτόνους

Οι τεχνικές αυτές επικεντρώνονται στην μελέτη κάθε κόμβου του δικτύου ως μέρος μια γειτονιάς. Δίνουν ιδιαίτερη βαρύτητα στους γειτονικούς κόμβους και πιο συγκεκριμένα στο πλήθος τους και όχι τόσο στην συνολική δομή του γράφου.

Κοινοί γείτονες (Common Neighbors (CN)) Ο συντελεστής αυτός είναι ίσως ο πιο διαδεδομένος στα κοινωνικά δίκτυα λόγω της απλότητας και της εύκολης κατανόησης του, ο οποίος δηλώνει πως η ομοιότητα δυο κόμβων εξαρτάται από τον αριθμό των κοινών τους γειτόνων [Newm01]. Για

παράδειγμα, σε ένα μέσο κοινωνικής δικτύωσης όπως το Facebook, αν δυο χρήστες έχουν πολλούς κοινούς φίλους τότε είναι πολύ πιθανόν οι δυο συγκεκριμένοι χρήστες να είναι και φίλοι μεταξύ τους. Μεταξύ δύο οποιονδήποτε κόμβων x και y του δικτύου, η μετρική των κοινών γειτόνων υπολογίζεται σύμφωνα με την Εξίσωση 2.1 παρακάτω. Είναι προφανές ότι όσο μεγαλύτερη είναι η τιμή της, τόσο πιο πιθανό γίνεται οι εν λόγω κόμβοι να συνδεθούν κάποια στιγμή στο μέλλον.

$$CN(x, y) = |\Gamma(x) \cap \Gamma(y)| \tag{2.1}$$

Συντελεστής Jaccard (Jaccard Coefficient (JC)) Ο συντελεστής Jaccard αποτελεί μια κανονικοποιημένη εκδοχή του συντελεστή των κοινών γειτόνων. Ο συντελεστής εξετάζει πόσους κοινούς γείτονες έχουν δυο κόμβοι και σε σχέση με τον συνολικό αριθμό γειτόνων των δυο κόμβων [Salt86] (Εξίσωση 2.2). Λαμβάνει τιμές στο διάστημα $[0, 1]$ και όπως και στην προηγούμενη περίπτωση, όσο μεγαλύτερη είναι η τιμή του, τόσο πιθανότερο είναι στο μέλλον οι δύο συγκεκριμένοι κόμβοι να συνδεθούν με κάποια ακμή.

$$JC(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{2.2}$$

Συντελεστής Adamic-Adar (Adamic-Adar Coefficient (AA)) Ο συγκεκριμένος συντελεστής προτάθηκε από τους Adamic και Adar για τον υπολογισμό της ομοιότητας μεταξύ δυο ιστοσελίδων [Adam01]. Μια παραλλαγή του μπορεί να χρησιμοποιηθεί και στα κοινωνικά δίκτυα, υπολογίζοντας την ομοιότητα μεταξύ δυο κόμβων. Ο συντελεστής εξετάζει τους κοινούς γειτόνους των δυο κόμβων και δίνει μεγαλύτερη βαρύτητα στους πιο «μοναχικούς» κοινούς γείτονες, δηλαδή τους κοινούς γείτονες που οι ίδιοι έχουν λίγους γειτόνους (Εξίσωση 2.3).

$$AA(x, y) = \sum_{z \in |\Gamma(x) \cap \Gamma(y)|} \frac{1}{\log |\Gamma(x)|} \tag{2.3}$$

Προτιμώμενη προσκόλληση (Preferential Attachment (PA)) Ο συντελεστής αυτός δηλώνει ότι οι πιο δημοφιλείς κόμβοι, δηλαδή οι κόμβοι με τους περισσότερους γειτόνους, έχουν μεγαλύτερες πιθανότητες να δημιουργήσουν νέες ακμές [Bara02] (Εξίσωση 2.4).

$$PA(x, y) = |\Gamma(x)| * |\Gamma(y)| \tag{2.4}$$

Δείκτης Sorensen (Sorensen Index (SI)) Ο δείκτης αυτός υπολογίζει τον αριθμό των κοινών γειτόνων μεταξύ δυο κόμβων και δηλώνει ότι κόμβοι με χαμηλό βαθμό θα έχουν υψηλότερο βαθμό ομοιότητας [Sore48] (Εξίσωση 2.5)

$$SI(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x)| + |\Gamma(y)|} \tag{2.5}$$

Δείκτης ομοιότητας Salton Cosine (Salton Cosine Similarity (SC)) Ο δείκτης αυτός προτάθηκε για τον υπολογισμό της ομοιότητας δυο διανυσμάτων, καθώς υπολογίζει το συνημίτονο της γωνίας δυο διανυσμάτων με βάση το εσωτερικό γινόμενο τους. Μια παραλλαγή του δείκτη μετρά τον βαθμό ομοιότητας δυο κόμβων ενός γράφου, όπως φαίνεται στην Εξίσωση 2.6.

$$SC(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{|\Gamma(x)| * |\Gamma(y)|}} \tag{2.6}$$

Δείκτης Hub Promoted (HP) Ο δείκτης αυτός προτάθηκε για τον υπολογισμό του βαθμού τοπολογικής επικάλυψης μεταξύ ζευγαριών υποστρωμάτων σε μεταβολικά δίκτυα [Rava02]. Μια παραλλαγή

του δείκτη για τον υπολογισμό της ομοιότητας μεταξύ δυο κόμβων ενός κοινωνικού δικτύου μπορεί να φανεί στην Εξίσωση 2.7.

$$HP(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{\min(|\Gamma(x)|, |\Gamma(y)|)} \tag{2.7}$$

Ακμές που προσπίπτουν σε δημοφιλείς κόμβους (κόμβους με πολύ μεγάλο βαθμό) πιθανόν να έχουν και μεγάλη τιμή από τον Δείκτη Hub Promoted, αφού ο παρανομαστής εξαρτάται από τον κόμβο με τον χαμηλότερο βαθμό.

Δείκτης Hub Depressed (HD) Ανάλογα με τον παραπάνω ορίζεται και αυτός ο δείκτης, ο οποίος έχει αντίστροφο αποτέλεσμα για τις προσπίπτουσες σε δημοφιλείς κόμβους ακμές [Zhou09] (Εξίσωση 2.8).

$$HD(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{\max(|\Gamma(x)|, |\Gamma(y)|)} \tag{2.8}$$

Συντελεστής Leicht-Holme-Newman (LHN) Ο συγκεκριμένος συντελεστής δίνει έμφαση σε κόμβους του δικτύου που έχουν αρκετούς κοινούς γειτόνους και μικρό βαθμό [Leic06] (Εξίσωση 2.9)

$$LHN(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x)| * |\Gamma(y)|} \tag{2.9}$$

Συντελεστής εξαρτώμενος από παράμετρο (Parameter-Dependent (PD)) Ένας πιο γενικός συντελεστής που αυξάνει τις σωστές προβλέψεις δημοφιλών και μη ακμών προτάθηκε από [Zhu12]. Ο τύπος του φαίνεται στην Εξίσωση 2.10.

$$PD(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{(|\Gamma(x)| * |\Gamma(y)|)^{\lambda}} \tag{2.10}$$

Το λ είναι μια ελεύθερη παράμετρος. Όταν $\lambda = 0$ ο τύπος μας δίνει τον συντελεστή κοινών γειτόνων, όταν $\lambda = 0,5$ ο τύπος μας δίνει τον δείκτη ομοιότητας Salton Cosine, ενώ τέλος όταν $\lambda = 1$ τον συντελεστή Leicht-Holme-Newman.

Κατανομή πόρων (Resource allocation (RA)) Ο δείκτης αυτός είναι παρόμοιος με τον Adamic-Adar, με τη διαφορά ότι "τιμωρούνται" περισσότεροι οι κόμβοι που έχουν υψηλότερο βαθμό (σε σύγκριση με τον AA) [Zhou09] (Εξίσωση 2.11).

$$RA(x,y) = \sum_{z \in |\Gamma(x) \cap \Gamma(y)|} \frac{1}{\Gamma(z)} \tag{2.11}$$

Σε αυτό το σημείο, πρέπει να τονιστεί ότι ο δείκτης Adamic-Adar και ο δείκτης RA δεν επικεντρώνονται στους γείτονες των υπό εξέταση κόμβων x και y, αλλά στους γείτονες των γειτόνων αυτών σε αντίθεση με κάθε άλλη μετρική που μελετήσαμε σε αυτήν την Ενότητα.

Ο Πίνακας 2.1 συνοψίζει τη χρονική πολυπλοκότητα των συντελεστών και των δεικτών που παρουσιάστηκαν προηγουμένως [Wang14]. Αν υποθέσουμε ότι n είναι ο μέσος αριθμός γειτόνων των κόμβων ενός γράφου, τότε παρατηρείται ότι οι δείκτες $CN, SI, SC, HP, HD, LHN, PD$ έχουν χρονική πολυπλοκότητα $\mathcal{O}(n^2)$, καθώς υπολογίζουν την τομή δυο συνόλων με τους γείτονες, η μετρική PA έχει πολυπλοκότητα $\mathcal{O}(2n)$, καθώς υπολογίζει τον αριθμό των γειτόνων μόνο δυο κόμβων, ενώ τέλος οι μετρικές AA και RA έχουν πολυπλοκότητα $\mathcal{O}(2n^2)$, αφού εξετάζουν την τομή δυο συνόλων για να βρουν τους κοινούς γείτονες αλλά και τους γείτονες των κοινών γειτόνων.

Συντελεστής	Χρονική πολυπλοκότητα
Κοινοί Γείτονες (CN)	$\mathcal{O}(n^2)$
Συντελεστής Jaccard (JC)	$\mathcal{O}(2n^2)$
Δείκτης Sorensen (SI)	$\mathcal{O}(n^2)$
Δείκτης ομοιότητας Salton Cosine (SC)	$\mathcal{O}(n^2)$
Δείκτης Hub Promoted (HP)	$\mathcal{O}(n^2)$
Δείκτης Hub Depressed (HD)	$\mathcal{O}(n^2)$
Συντελεστής Leicht-Holme-Newman (LHN)	$\mathcal{O}(n^2)$
Συντελεστής εξαρτώμενος από παράμετρο (PD)	$\mathcal{O}(n^2)$
Συντελεστής Adamic-Adar (AA)	$\mathcal{O}(2n^2)$
Προτιμώμενη προσκόλληση (PA)	$\mathcal{O}(2n)$
Κατανομή πόρων (RA)	$\mathcal{O}(2n^2)$

Πίνακας 2.1: Χρονική πολυπλοκότητα μετρικών βασισμένων στην τοπολογία του γράφου

2.2.2 Τεχνικές βασισμένες στη διαδρομή

Στην παραπάνω Ενότητα μελετήθηκαν συντελεστές που υπολογίζουν τον βαθμό ομοιότητας δυο κόμβων με βάση τους γειτόνους τους. Αυτή η πρακτική είναι πολύ περιορισμένη, καθώς δεν λαμβάνει υπόψιν το υπόλοιπο δίκτυο. Σε αυτήν την Ενότητα θα μελετηθούν τεχνικές που βασίζονται στις *διαδρομές* (paths) μεταξύ δυο κόμβων, προσφέροντας μια γενικότερη λύση στο πρόβλημα της πρόβλεψης ακμών, καθώς εξερευνούν όλο τον γράφο και δεν περιορίζονται μόνο στους γειτονικούς κόμβους.

Βασική παράμετρος στην μελέτη των τεχνικών αυτής της Ενότητας είναι ο ορισμός του πίνακα γειτνίασης ενός γράφου. Ο *πίνακας γειτνίασης* (adjacency matrix) είναι τετραγωνικός και αποτελεί μια άλλη μορφή αναπαράστασης του γράφου ενός δικτύου. Το στοιχείο (i, j) του πίνακα παίρνει τιμή 0 ή 1 και δηλώνει την ύπαρξη ακμής μεταξύ του κόμβου i και j (τιμή 1) ή την έλλειψη της (τιμή 0). Η χρήση του πίνακα γειτνίασης είναι απαραίτητη για την αλγεβρική μελέτη των γράφων και πιο συγκεκριμένα τον υπολογισμό των μονοπατιών μεταξύ δυο κόμβων. Γενικά, ο αριθμός των διαδρομών μήκους l μεταξύ του κόμβου i και j, είναι το στοιχείο (i, j) του πίνακα A^l [Bigg93].

Τοπικό μονοπάτι (Local Path (LP)) Ο δείκτης αυτός εξετάζει τα μονοπάτια μήκους 2 και 3 που υπάρχουν μεταξύ δυο κόμβων για να βρει τον βαθμό ομοιότητας μεταξύ τους [Lu09] (Εξίσωση 2.12).

$$LP(x, y) = A^2 + \alpha A^3 \tag{2.12}$$

Ο συντελεστής α παίρνει τιμές κοντά στο 0, δίνοντας μεγαλύτερη βαρύτητα στα μονοπάτια μήκους 2.

Δείκτης Katz Γενίκευση του παραπάνω δείκτη αποτελεί ο Δείκτης Katz, ο οποίος λαμβάνει υπόψη όλα τα μονοπάτια οποιουδήποτε μήκους μεταξύ δυο κόμβων για να βρει τον βαθμό ομοιότητας τους [Katz53] (Εξίσωση 2.13).

$$Katz(x, y) = \sum_{l=1}^{\infty} \beta^l \left|\text{path}_{x,y}^l\right| = \beta A + \beta^2 A^2 + \beta^3 A^3 + \dots \tag{2.13}$$

Η μεταβλητή l δηλώνει το μήκος του μονοπατιού. Ο συντελεστής β είναι σχετικά μικρός επιβραβεύοντας τις διαδρομές με μικρό μήκος και επιβαρύνοντας τις διαδρομές με μεγάλο μήκος. Το σύνολο $\left|\text{path}_{x,y}^l\right|$ περιέχει όλα τα μονοπάτια μήκους l που υπάρχουν μεταξύ των κόμβων x και y.

Είναι προφανές ότι για τιμές του συντελεστή β κοντά στο μηδέν, όσο αυξάνει το l τόσο πιο μικρή συνεισφορά έχει ο όρος $\beta^l A^l$, ενώ παράλληλα τόσο μεγαλώνει και η πολυπλοκότητα του υπολογισμού του πίνακα A^l. Αυτό σημαίνει πως ο Δείκτης Katz δίνει έμφαση στα σχετικά μικρά μονοπάτια μεταξύ δυο κόμβων.

Ο δείκτης Katz μπορεί να χρησιμοποιηθεί σε γράφους με βάρη στις ακμές, αλλά και σε γράφους χωρίς βάρη. Ας σκεφτούμε για παράδειγμα έναν γράφο μιας ακαδημαϊκής κοινότητας, όπου οι κόμβοι του αντιστοιχούν στους καθηγητές / ερευνητές και οι ακμές του δηλώνουν συνεργασία που έχει επιτευχθεί μεταξύ των δυο κόμβων / ερευνητών. Αν ο γράφος δεν έχει βάρη, τότε όλες οι ακμές θα έχουν τιμή 1, ή αλλιώς $\left|\text{path}_{x,y}^{1}\right| = 1$. Αυτό σημαίνει πως αν υπάρχει ακμή μεταξύ δυο οποιονδήποτε κόμβων, δηλαδή αν υπάρχει μονοπάτι μήκους 1 μεταξύ τους, αυτό θα είναι μοναδικό. Αντίθετα, σε γράφο με βάρη θα ίσχυε $\left|\text{path}_{x,y}^{1}\right| = k$ με το k να δηλώνει τις φορές που δυο ερευνητές x και y έχουν συνεργαστεί.

Στην περίπτωση του γράφου χωρίς βάρη ο δείκτης έχει όνομα *Katz χωρίς βάρη* (Katz unweighted), ενώ αλλιώς ονομάζεται δείκτης *Katz με βάρη* (Katz weighted).

Relation Strength Similarity (RSS) Ο δείκτης RSS χρησιμοποιείται στους γράφους με βάρη στις ακμές. Στηρίζεται στον δείκτη ομοιότητας $R(x, y)$ μεταξύ δυο κόμβων που ισοδυναμεί και με το βάρος της ακμής μεταξύ των δυο συγκεκριμένων γειτονικών κόμβων [Chen12]. Ας υποθέσουμε ότι υπάρχουν L μονοπάτια $p_1, p_2, \ldots, p_L$ μήκους μικρότερου από r μεταξύ των κόμβων x και y. Επίσης, το μονοπάτι p_l σχηματίζεται από K κόμβους $z_1, z_2, \ldots, z_k$ με $K < r$. Τότε, ο δείκτης RSS ορίζεται όπως παρακάτω.

$$RSS(x, y) = \sum_{l=1}^{L} R_{p_l}^{*}(x, y) \tag{2.14}$$

$$R_{p_l}^{*}(x, y) = \begin{cases} \prod_{k=1}^{K} R(z_k, z_{k+1}) & \text{για } K \leq r \\ 0 & \text{διαφορετικά} \end{cases} \tag{2.15}$$

Δείκτης Friendlink (FL) Ο δείκτης Friendlink υπολογίζει την ομοιότητα μεταξύ δυο κόμβων, εξετάζοντας όλα τα μονοπάτια συγκεκριμένου μήκους μεταξύ τους [Papa12]. Η συνεισφορά κάθε μονοπατιού στον βαθμό ομοιότητας, είναι αντιστρόφως ανάλογη του μήκους του μονοπατιού (Εξίσωση 2.16).

$$FL(x, y) = \sum_{i=1}^{l} \frac{1}{i-1} \frac{\left|\text{paths}_{x,y}^{i}\right|}{\prod_{j=2}^{i} n - j} \tag{2.16}$$

όπου, n είναι ο αριθμός των κόμβων του δικτύου, l είναι το μέγιστο εξεταζόμενο μήκος μονοπατιού και $\text{paths}_{x,y}^{i}$ είναι το σύνολο των μονοπατιών μήκους i μεταξύ των κόμβων x και y.

2.2.3 Τεχνικές βασισμένες στους τυχαίους περιπάτους

Μέχρι τώρα έχουν μελετηθεί τεχνικές που μετρούν την ομοιότητα μεταξύ δυο κόμβων με βάση τους γειτονικούς τους κόμβους αλλά και τα διάφορα μονοπάτια που μπορεί να υπάρχουν μεταξύ τους. Σε αυτή την Ενότητα θα γίνει αναφορά στους τυχαίους περιπάτους.

Ένας *τυχαίος περίπατος* (random walk) σε έναν γράφο G είναι μια τυχαία ακολουθία κόμβων $(u_0, u_1, ...)$, ξεκινώντας από αρχικό κόμβο u_0 και επιλέγοντας τυχαία τον επόμενο κόμβο u_1 του περιπάτου με βάση κατανομή πιθανότητας σ_0. Στην συνέχεια, επιλέγεται τυχαία γειτονικός κόμβος u_2 με βάση κατανομή πιθανότητας σ_1, ώστε ο περίπατος να συνεχιστεί [Lova96].

Ο [Lova96] μελέτησε τους τυχαίους περιπάτους με χρήση των πινάκων των γράφων και τις ιδιοτιμές αυτών. Πιο συγκεκριμένα, μελέτησε την σημαντικότητα πινάκων που ορίζονται από την δομή του γράφου και απέδειξε τον ρόλο που παίζουν οι ιδιοτιμές και τα ιδιοδιανύσματα ορισμένων πινάκων στην μελέτη των τυχαίων περιπάτων.

Αρχικά, ορίζεται ο διαγώνιος πίνακας D_A διαστάσεων nxn που περιέχει τον βαθμό των n κόμβων του γράφου. Με την βοήθεια του πίνακα D_A, μπορεί να οριστεί ο *πίνακας μεταβάσεων* (transition matrix) $M = D_A A$ του γράφου, όπου κάθε στοιχείο του $M_{x,j}$ δηλώνει την πιθανότητα μετάβασης από τον κόμβο x στον κόμβο j.

Πολύ σημαντική παράμετρος στην μελέτη των τυχαίων περιπάτων αποτελεί ο ορισμός της *στατικής πιθανότητας* (stationary probability). Η στατική πιθανότητα είναι χαρακτηριστικό κάθε κόμβου του δικτύου και δηλώνει έναν βαθμό σημαντικότητάς του, ενώ μπορεί να υπολογιστεί αναδρομικά (Εξίσωση 2.17).

$$P_{t+1} = M^\top P_t \tag{2.17}$$

Το διάνυσμα P_t έχει διαστάσεις $nx1$, όπου n ο αριθμός των κόμβων του γράφου, και περιέχει τις στατικές πιθανότητες όλων των κόμβων του γράφου σε κάθε βήμα της αναδρομής. Το διάνυσμα συγκλίνει σε οριακές τιμές σχετικά γρήγορα και η αναδρομή σταματά όταν η απόσταση μεταξύ των διανυσμάτων P_t και P_{t+1} είναι ελάχιστη.

Τέλος, ορίζεται ο πίνακας N ο οποίος αποτελεί την συμμετρική μορφή του πίνακα μεταβάσεων M και φαίνεται στην Εξίσωση 2.18.

$$N = D^{\frac{1}{2}} A D^{\frac{1}{2}} \tag{2.18}$$

Λόγω της συμμετρικότητας του, ο πίνακας N μπορεί να γραφτεί και με χρήση των ιδιοτιμών και των ιδιοδιανυσμάτων του όπως φαίνεται στην Εξίσωση 2.19.

$$N = \sum_{k=1}^{n} \lambda_k v_k v_k^\top \tag{2.19}$$

όπου λ_k είναι οι ιδιοτιμές του πίνακα N και u_k τα ιδιοδιανύσματα του με μοναδιαίο μήκος.

Ο ορισμός του πίνακα μεταβάσεων M, της στατικής πιθανότητας κάθε κόμβου και του συμμετρικού πίνακα N με τις ιδιοτιμές και τα ιδιοδιανύσματα του, αποτελούν βασικά συστατικά στην θεμελίωση πολλών από των μετρικών που θα παρουσιαστούν σε αυτήν την ενότητα.

Hitting Time (HT) Ο δείκτης Hitting Time είναι ο αναμενόμενος αριθμός βημάτων ενός τυχαίου περιπάτου από έναν κόμβο x σε έναν κόμβο y. Ο αναδρομικός τύπος του δείκτη παρουσιάζεται στην Εξίσωση 2.20 παρακάτω [Fous07].

$$HT(x, y) = 1 + \sum_{\omega \in \Gamma(\chi)} M^\top_{\chi\omega} HT(\omega, y) \tag{2.20}$$

Είναι προφανές ότι η πολυπλοκότητα του υπολογισμού της μετρικής με τον αναδρομικό τύπο 2.20 αυξάνει εκθετικά σε κάθε βήμα της αναδρομής και άρα, γενικά, ο υπολογισμός είναι αρκετά δύσκολος.

Ο ορισμός της μετρικής μπορεί να γίνει και με χρήση των ιδιοτιμών και ιδιοδιανυσμάτων του πίνακα N όπως φαίνεται στην Εξίσωση 2.21

$$HT(x, y) = 2n \sum_{k=2}^{n} \frac{1}{1 - \lambda_k} \left(\frac{v_{ky}^2}{d(y)} - \frac{v_{kx} v_{ky}}{\sqrt{d(x) d(y)}} \right) \tag{2.21}$$

όπου n συμβολίζει τον αριθμό κόμβων του γράφου και $d(x)$ τον βαθμό του κόμβου x.

Commute Time (CT) Ο δείκτης CT είναι πιο συμμετρικός σε σχέση με τον δείκτη HT, καθώς δεν υπολογίζει μόνο τον αναμενόμενο αριθμό βημάτων από έναν κόμβο x σε έναν κόμβο y, αλλά προσθέτει και τον αναμενόμενο αριθμό βημάτων από τον κόμβο y πίσω στον κόμβο x. Ο τύπος της μετρικής θα μπορούσε να δοθεί πολύ απλά.

$$CT(x, y) = HT(x, y) + HT(y, x) \tag{2.22}$$

Ο [Lova96] δίνει επίσης τον τύπο της μετρικής χρησιμοποιώντας πάλι τον πίνακα N με τις ιδιοτιμές και τα ιδιοδιανύσματα του.

$$CT(x, y) = 2m \sum_{k=2}^{n} \frac{1}{1 - \lambda_k} \left(\frac{v_{ky}^2}{d(y)} - \frac{v_{kx}}{d(x)} \right)^2 \tag{2.23}$$

με τα μεγέθη $\lambda_k, v_k, d(y)$ να είναι ίδια όπως ορίστηκαν στην εισαγωγή.

Cosine Similarity Time (CST) Για να ορίσουμε τον δείκτη αυτό, αρχικά πρέπει να ορίσουμε τον πίνακα L' που είναι ο ψευδοαντίστροφος πίνακας του πίνακα L που δίνεται από τον τύπο:

$$L = D_A - A \tag{2.24}$$

Ο τύπος της μετρικής CST δίνεται παρακάτω.

$$CST(x, y) = \frac{L'_{x,y}}{\sqrt{L'_{x,x}L'_{y,y}}} \tag{2.25}$$

SimRank Ο δείκτης SimRank ορίζεται με αναδρομικό τρόπο και στηρίζεται στην ιδέα ότι δυο κόμβοι έχουν ομοιότητα αν συνδέονται σε όμοιους κόμβους [Jeh02] (Εξίσωση 2.26)

$$\text{SimRank}(x, y) = \begin{cases} 1 & \text{, όταν } x = y \\ \gamma \frac{\sum_{\alpha \in \Gamma(x)} \sum_{\beta \in \Gamma(y)} \text{SimRank}_{(a,b)}}{|\Gamma(x)||\Gamma(y)|} & \text{, διαφορετικά} \end{cases} \tag{2.26}$$

Ο συντελεστής γ παίρνει τιμές μεταξύ του 0 και του 1 και δίνει μεγαλύτερη βαρύτητα στους κόμβους που βρίσκονται κοντά στους αρχικούς κόμβους x και y. Καθώς απομακρυνόμαστε από αυτούς η σημαντικότητα της ομοιότητας των κόμβων μειώνεται.

Η μετρική SimRank μπορεί να εξηγηθεί με όρους των τυχαίων περιπάτων, αν θεωρήσουμε ότι δυο περιπατητές ξεκινούν από τυχαίους κόμβους x και y του γράφου. Η μετρική δηλώνει πόσο σύντομα οι δυο περιπατητές θα βρεθούν στον ίδιο κόμβο, αν περιδιαβαίνουν τυχαίες ακμές του γράφου.

Λόγω της αναδρομικής δομής της μετρικής, η χρονική πολυπλοκότητά της είναι απαγορευτική για χρήση σε μεγάλα κοινωνικά δίκτυα.

Κανονικοποιημένο Hitting Time (Normalized Hitting Time (NHT)) Ένα χαρακτηριστικό των γράφων είναι ότι οι δημοφιλείς κόμβοι, δηλαδή κόμβοι με πολλούς γειτόνους, θα είναι προσπελάσιμοι πολύ πιο εύκολα σε σύγκριση με κόμβους με χαμηλό βαθμό. Με άλλα λόγια, αναμένουμε ότι ο δείκτης Hitting Time για δημοφιλείς κόμβους θα έχει σχετικά μικρή τιμή, δηλαδή ο αναμενόμενος αριθμός βημάτων για να φτάσει κάποιος σε έναν δημοφιλή κόμβο θα είναι μικρός. Για να αντιμετωπιστεί αυτή η δυσκολία, χρησιμοποιείται η πιθανότητα να βρεθεί ο τυχαίος περίπατος στον κόμβο y στην στατική κατανομή (Εξίσωση 2.17). Αν πολλαπλασιαστεί με την μετρική Hitting Time δίνει μια κανονικοποιημένη μορφή της μετρικής όπως φαίνεται από την παρακάτω Εξίσωση.

$$NHT(x, y) = p_y * HT(x, y) \tag{2.27}$$

Κανονικοποιημένο Commute Time (Normalized Commute Time (NCT)) Με παρόμοιο τρόπο μπορούμε να ορίσουμε και την κανονικοποιημένη μορφή του δείκτη Commute Time, χρησιμοποιώντας την πιθανότητα να βρεθεί ο τυχαίος περίπατος στους κόμβους x και y στην στατική κατανομή. Ο τύπος του νέου δείκτη φαίνεται παρακάτω.

$$NCT(x, y) = p_y * HT(x, y) + p_x * HT(y, x) \tag{2.28}$$

Rooted PageRank (RPR) Ένα πρόβλημα της μετρικής Hitting Time είναι ότι πολλές φορές ο τυχαίος περίπατος μπορεί να ακολουθήσει μονοπάτια πολύ απομακρυσμένα. Για παράδειγμα, μπορεί ο κόμβος x να είναι πάρα πολύ κοντά τοπολογικά στον κόμβο y, αλλά λόγω της τυχαίας επιλογής ακμών, ο τυχαίος περίπατος να οδηγηθεί σε ένα μέρος του γράφου πολύ μακριά από τους δυο κόμβους. Επίσης, υπάρχει η πιθανότητα ο τυχαίος περίπατος να εγκλωβιστεί σε αποκομμένες γειτονιές / περιοχές του γράφου, όπου η σύνδεση με τον υπόλοιπο γράφο γίνεται μόνο από μια ακμή. Για την επίλυση αυτού του προβλήματος δημιουργήθηκε ο δείκτης Rooted PageRank.

Η ιδέα είναι ίδια με τους τυχαίους περιπάτους στον δείκτη Hitting Time, αλλά αυτή τη φορά υπάρχει πιθανότητα σε κάθε βήμα του αλγορίθμου, ο τυχαίος περίπατος να επανεκκινήσει από την αρχή.

Η μετρική Rooted PageRank είναι η στατική πιθανότητα να βρεθούμε στον κόμβο y αν ξεκινήσουμε από το κόμβο x και σε κάθε βήμα του τυχαίου περιπάτου, υπάρχει πιθανότητα α να επιστρέψουμε στον κόμβο x και να ξεκινήσουμε από την αρχή και πιθανότητα $1-\alpha$ να προχωρήσουμε στον επόμενο τυχαίο γειτονικό κόμβο [Libe03].

Όπως και στο κανονικοποιημένο Hitting Time, μπορεί να υπολογιστεί η στατική πιθανότητα κάθε κόμβου με αναδρομικό τρόπο. Ο τύπος είναι διαφορετικός, καθώς πρέπει να εισαχθεί και η πιθανότητα επανεκκίνησης α (Εξίσωση 2.29).

$$P_{t+1} = \alpha M^{\top} P_t + \frac{1-\alpha}{n} I \tag{2.29}$$

όπου n ο αριθμός των κόμβων του γράφου και I ο μοναδιαίος πίνακας.

Επίσης, οι [Wang14] ορίζουν την μετρική Rooted PageRank με τον παρακάτω τύπο.

$$RPR = (1-\alpha)(I - \alpha D^{-1} A)^{-1} \tag{2.30}$$

όπου οι πίνακες D και A είναι όπως ορίστηκαν παραπάνω.

Με τον αλγόριθμο PageRank γίνεται αντιληπτό ότι μέρη του γράφου που βρίσκονται μακριά από τον αρχικό κόμβο δεν εξερευνώνται σχεδόν ποτέ.

PropFlow (PF) Ο δείκτης PropFlow είναι παρόμοιος με το Rooted PageRank με την μόνη διαφορά ότι ο τυχαίος περίπατος που ξεκινά από τον κόμβο x και λήγει στον κόμβο y δεν έχει μήκος μεγαλύτερο από l. Ο περιορισμένος τυχαίος περίπατος επιλέγει κόμβους για να κινηθεί, με βάση το βάρος των ακμών και τερματίζει όταν συναντήσει το y ή όταν επισκεφτεί οποιονδήποτε άλλο κόμβο δεύτερη φορά [Lich10].

Αν οι τυχαίοι κόμβοι x και y συνδέονται άμεσα, τότε ο ορισμός του PropFlow δίνεται από τον παρακάτω τύπο.

$$PF(x,y) = PF(a,x)\frac{w_{xy}}{\sum\limits_{k \in \Gamma(x)} w_{xk}} \tag{2.31}$$

όπου k είναι γείτονας του κόμβου x με βάθος μεγαλύτερο από το βάθος του κόμβου x από τον αρχικό κόμβο, w_{xy} είναι το βάρος της ακμής μεταξύ των κόμβων x και y και α είναι ο προηγούμενος κόμβος του x σε ένα τυχαίο μονοπάτι. Αν ο τυχαίος κόμβος x είναι ο αρχικός κόμβος, τότε $PF(\alpha, x) = 1$.

Αν οι τυχαίοι κόμβοι x και y δεν συνδέονται άμεσα, τότε $PF(x,y)$ είναι το άθροισμα της μετρικής καθώς διασχίζει όλα τα συντομότερα μονοπάτια από το x στο y.

Σε αντίθεση με το Rooted PageRank, ο δείκτης PropFlow δεν χρησιμοποιεί επανεκκινήσεις των τυχαίων περιπάτων, αλλά χρησιμοποιεί μια παραλλαγή της αναζήτησης κατά βάθος σε γράφο με περιορισμό στο βάθος, το όριο l. Εξαιτίας αυτής της διαφοροποίησης, ο PropFlow είναι πολύ πιο εύκολη στον υπολογισμό της από τις μετρικές Rooted PageRank και SimRank.

2.3 Τεχνικές βασισμένες στην κοινωνική θεωρία

Τα τελευταία χρόνια όλο και περισσότερες εργασίες πάνω στο πρόβλημα της πρόβλεψης ακμών στα κοινωνικά δίκτυα, αναφέρονται σε τεχνικές πρόβλεψης που πηγάζουν από την *κοινωνική θεωρία* (social theory). Μετρικές όπως η τριαδική κλειστότητα, οι ισχυροί και ασθενείς δεσμοί [Liu13] και η ομοφυλία [Yang11], μπορούν να βελτιώσουν την απόδοση στην πρόβλεψη των ακμών, χρησιμοποιώντας χρήσιμες πληροφορίες για την κοινωνική δραστηριότητα χρηστών του δικτύου. Η μελέτη τέτοιων τεχνικών, ωστόσο, ξεφεύγει από τους σκοπούς της συγκεκριμένης εργασίας και για το λόγο αυτό δεν θα αναλυθούν περισσότερο.

Κεφάλαιο 3

Ευφυείς τεχνικές πρόβλεψης ακμών στα μέσα κοινωνικής δικτύωσης

3.1 Εισαγωγή

Στο προηγούμενο Κεφάλαιο μελετήθηκαν ορισμένες απλές, γραφοθεωρητικές τεχνικές επίλυσης του προβλήματος της πρόβλεψης ακμών στα μέσα κοινωνικής δικτύωσης. Αυτές οι τεχνικές στηρίζονται στην μελέτη των χαρακτηριστικών των κόμβων, εάν υπάρχουν, αλλά και στην τοπολογία του γράφου γενικότερα. Εκτός των προαναφερόμενων τεχνικών όμως, που στην πλειοψηφία τους είναι ντετερμινιστικές, το πρόβλημα της πρόβλεψης ακμών μπορεί να μελετηθεί και από την σκοπιά των *ευφυών τεχνικών* (intelligent techniques), με την κατάλληλη διατύπωσή του ως ένα πρόβλημα *δυαδικής ταξινόμησης* (binary classification).

Η δυαδική ταξινόμηση είναι η διαδικασία τοποθέτησης των στοιχείων ενός συνόλου σε δυο προκαθορισμένες κατηγορίες με βάση κάποιους κανόνες. Ένα παράδειγμα δυαδικής ταξινόμησης είναι η εξέταση ενός ασθενούς από ένα ακτινολογικό κέντρο με σκοπό να προσδιοριστεί αν έχει καρκίνο ή όχι. Στην περίπτωση που εξετάζεται σε αυτήν την εργασία, το πρόβλημα δυαδικής ταξινόμησης μπορεί να οριστεί με τον παρακάτω τρόπο.

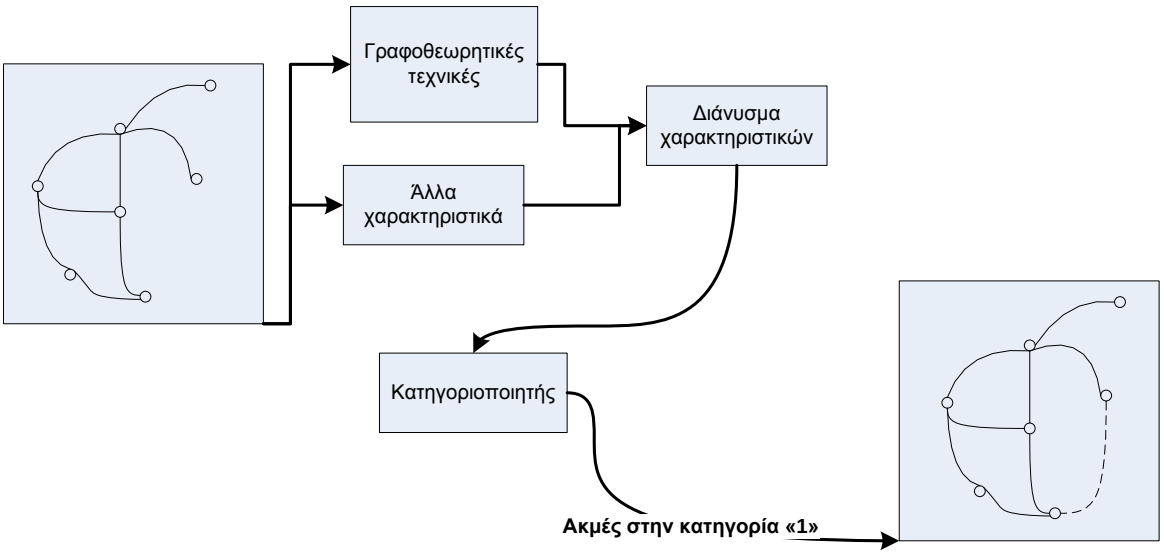
Έστω x και y δυο κόμβοι ενός γράφου $G(V, E)$ και $l^{(x,y)}$ η ετικέτα που δείχνει σε ποια κατηγορία ανήκει η ακμή (x, y) . Οι δυνατές κατηγορίες είναι δύο: είτε να υπάρχει ακμή μεταξύ των δύο κόμβων είτε όχι. Με βάση τις δυο αυτές κατηγορίες, η ετικέτα μπορεί να πάρει τις ακόλουθες τιμές.

$$l^{(x,y)} = \begin{cases} 1, \text{αν } (x, y) \in E \\ 0, \text{αν } (x, y) \notin E \end{cases} \tag{3.1}$$

Έχοντας ορίσει το δυαδικό ισοδύναμο πρόβλημα πρόβλεψης ακμών, μπορούν πλέον να χρησιμοποιηθούν πολλά *μοντέλα επιβλεπομένης ταξινόμησης* (supervised learning models). Στο παρόν κεφάλαιο θα παρουσιαστούν διάφορα μοντέλα ταξινόμησης, όπως ο *απλός μπεϋζιανός ταξινομητής* (naive bayesian classifier), οι k-*πλησιέστεροι γείτονες* (k-nearest neighbors), τα *δέντρα αποφάσεων* (decision trees), οι *μηχανές διανυσμάτων υποστήριξης* (support vector machines) καθώς και η μέθοδος της *στοχαστικής καθόδου κλίσης* (stochastic gradient descent).

Η πιο γνωστή τεχνική ταξινόμησης που μπορεί να χρησιμοποιηθεί για το πρόβλημα της πρόβλεψης ακμών είναι η μέθοδος με χρήση ενός *διανύσματος χαρακτηριστικών* (feature-based classification). Κάθε ζεύγος κόμβων, είτε συνδέονται μεταξύ τους με ακμή είτε όχι, περιγράφεται από ένα διάνυσμα χαρακτηριστικών καθώς και από την ετικέτα της κατηγορίας στην οποία ανήκει. Η επιλογή των σωστών χαρακτηριστικών είναι πολύ μεγάλης σημασίας για την ορθή απόδοση του ταξινομητή. Πριν προχωρήσουμε στην ανάλυση των χαρακτηριστικών, το Σχήμα 3.1 δίνει μια γενικότερη εικόνα του δυαδικού προβλήματος πρόβλεψης ακμών και του πώς αυτό μπορεί να επιλυθεί.

Το πρώτο βήμα είναι η δημιουργία του διανύσματος των χαρακτηριστικών για κάθε ζεύγος κόμβων του δικτύου, είτε αυτοί συνδέονται μεταξύ τους με ακμή είτε όχι. Στην συνέχεια, τα διανύσματα αυτά μαζί με τις αντίστοιχες ετικέτες τους αποτελούν την είσοδο του εκάστοτε ταξινομητή. Ο ταξινομητής εκπαιδεύεται με βάση τα στιγμιότυπα που έλαβε ως είσοδο και πλέον λαμβάνοντας ένα νέο διάνυσμα χαρακτηριστικών μπορεί να αποφασίσει σε ποια κατηγορία πρέπει να τοποθετηθεί. Τα



Σχήμα 3.1: Διαδικασία πρόβλεψης ακμών με χρήση ευφυών τεχνικών

Τύπος	**Χαρακτηριστικό**
Κόμβος	Βαθμός
Δίκτυο	Αριθμός κόμβων, αριθμός ακμών, βαθμός συσταδοποίησης, μέσος βαθμός κόμβων κτλ
Τοπολογία δικτύου	Κοινοί γείτονες, συντελεστής Jaccard, συντελεστής Adamic/Adar, Προτιμώμενη Προσκόλληση, Τοπικά μονοπάτια, μετρική Katz, SimRank, μετρική Hitting Time κτλ
Μη-τοπολογικά	Όνομα συγγραφέων, τίτλος κειμένου, ημερομηνία έκδοσης, αριθμός βασικών λέξεων (keywords) κτλ

Πίνακας 3.1: Πιθανά χαρακτηριστικά χρήσιμα για την δημιουργία του διανύσματος εισόδου του ταξινομητή

ζεύγη κόμβων που κατηγοριοποιούνται στην κλάση ”1” είναι πιθανό να συνδεθούν μεταξύ τους με ακμή σε κάποιο μελλοντικό στιγμιότυπο του γράφου.

Το πιο δύσκολο τμήμα της παραπάνω διαδικασίας είναι η σωστή επιλογή των χαρακτηριστικών για την δημιουργία του διανύσματος χαρακτηριστικών κάθε ζεύγους κόμβων. Ορισμένα χαρακτηριστικά μπορεί να εξαχθούν απευθείας από τον γράφο, όπως για παράδειγμα, ο βαθμός των κόμβων που αποτελούν το ζεύγος. Επίσης, τα χαρακτηριστικά μπορούν να προέλθουν και από τις γραφοθεωρητικές τεχνικές που μελετήθηκαν στο Κεφάλαιο 2.

Γενικότερα στοιχεία του γράφου θα μπορούσαν επίσης να αποτελέσουν σημαντικά χαρακτηριστικά, όπως ο αριθμός των ακμών, ο αριθμός των κόμβων και ο μέσος βαθμός των κόμβων. Οι [Wang14] δίνουν έναν πίνακα με τα πιθανά χαρακτηριστικά που θα μπορούσαν να συνθέσουν το διάνυσμα εισόδου για τον ταξινομητή, αναφερόμενοι στο πρόβλημα πρόβλεψης ακμών σε ένα δίκτυο συγγραφέων, όπου οι κόμβοι του δικτύου είναι οι συγγραφείς και οι ακμές μεταξύ κόμβων αποτελούν κείμενα που γράφτηκαν από δυο συγγραφείς / κόμβους του δικτύου (Πίνακας 3.1).

Είναι χρήσιμο να ειπωθεί ότι τα μη-τοπολογικά χαρακτηριστικά συναρτώνται άμεσα με το υπό μελέτη κοινωνικό δίκτυο. Στο παράδειγμα του κοινωνικού δικτύου των συγγραφέων που αναφέρθηκε παραπάνω, ο τίτλος του κειμένου και η ημερομηνία έκδοσης έχουν σημασία. Είναι σαφές, ότι για άλλα κοινωνικά δίκτυα τέτοιου είδους πληροφορίες μπορεί να μην υπάρχουν ή μπορεί να είναι ελλιπώς συμπληρωμένες. Επίσης, η αναγνώριση και συλλογή τέτοιας πληροφορίας απαιτεί γνώση του αντίστοιχου γνωστικού αντικειμένου. Παρ’ όλα αυτά η χρήση μη-τοπολογικών χαρακτηριστικών μπορεί να βελτιώσει την απόδοση του ταξινομητή.

Αφού παρουσιάστηκαν τα χαρακτηριστικά που θα μπορούσαν να αποτελέσουν τη βάση του διανύσματος των χαρακτηριστικών για κάθε ζεύγος κόμβων, στο υπόλοιπο Κεφάλαιο θα εξεταστεί η χρήση διαφόρων ταξινομητών για την επίλυση του προβλήματος πρόβλεψης ακμών με χρήση ευφυών τεχνικών.

3.2 Απλός Μπεϋζιανός Ταξινομητής

Στην μηχανική μάθηση, ο *απλός μπεϋζιανός ταξινομητής* (naive bayesian classifier) είναι ένας απλός πιθανοτικός ταξινομητής που στηρίζεται στο θεώρημα του Bayes, το οποίο θεωρεί ότι τα χαρακτηριστικά του διανύσματος εισόδου είναι ανεξάρτητα μεταξύ τους [Zhan04a]. Ο μπεϋζιανός ταξινομητής χρησιμοποιείται για πολλά χρόνια με επιτυχία κυρίως ως μέθοδος *ταξινόμησης κειμένου* (text classifier). Χαρακτηριστικό παράδειγμα ταξινόμησης κειμένου είναι η κατηγοριοποίηση ενός εισερχομένου προσωπικού ηλεκτρονικού μηνύματος ως spam ή ως χρήσιμο μήνυμα.

Όπως ειπώθηκε και στην εισαγωγή του παρόντος Κεφαλαίου, σκοπός του μπεϋζιανού ταξινομητή και γενικότερα όλων των ταξινομητών, είναι να δώσει μια *ετικέτα* (label) σε κάθε νέο δείγμα του προβλήματος που λαμβάνει. Η ετικέτα θα δηλώνει σε ποια κατηγορία ανήκει το συγκεκριμένο δείγμα. Οι κατηγορίες για κάθε πρόβλημα κατηγοριοποίησης είναι συγκεκριμένες και καθορισμένες εκ των προτέρων.

Ο *μπεϋζιανός κανόνας* (Bayes' rule) δεν είναι ένας μοναδικός αλγόριθμος εκπαίδευσης ταξινομητών, αλλά μια οικογένεια αλγορίθμων που περιλαμβάνει πολλές παραλλαγές ταξινομητών. Η τιμή ενός χαρακτηριστικού (feature) που υπάρχει μέσα στο διάνυσμα χαρακτηριστικών, είναι ανεξάρτητη από κάθε άλλη τιμή χαρακτηριστικού στο διάνυσμα. Κάθε ένα χαρακτηριστικό του διανύσματος συνεισφέρει ανεξάρτητα ώστε το δείγμα τελικά να ανήκει σε μια συγκεκριμένη κατηγορία του προβλήματος.

Ας υποθέσουμε ότι μας δίνεται ένα δείγμα του προβλήματος προς ταξινόμηση, το οποίο αναπαρίσταται ως διάνυσμα χαρακτηριστικών

$$X = (x_1, \ldots, x_2) \tag{3.2}$$

Ο απλός μπεϋζιανός ταξινομητής δίνει σε αυτό το δείγμα πιθανότητες να ανήκει σε κάθε κατηγορία του προβλήματος ως εξής.

$$p(C_k|x_1, \ldots, x_n) \tag{3.3}$$

όπου k ο αριθμός των κατηγοριών του προβλήματος. Σύμφωνα με το θεώρημα του Bayes η παραπάνω πιθανότητα ισοδυναμεί με

$$p(C_k|X) = \frac{p(C_k)p(X|C_k)}{p(X)} \tag{3.4}$$

Υποθέτοντας ότι όντως τα χαρακτηριστικά του δείγματος είναι ανεξάρτητα μεταξύ τους, ο όρος $p(X|C_k)$ μπορεί να γραφεί όπως παρακάτω.

$$p(X|C_k) = p(x_1, \ldots, x_n|C_k) = p(x_1|C_k) * p(x_2|C_k) * \ldots * p(x_n|C_k) \tag{3.5}$$

που αποτυπώνει αυτό που αναφέρθηκε προηγουμένως, ότι δηλαδή η συνεισφορά του κάθε χαρακτηριστικού είναι ανεξάρτητη της κατηγορίας που ανήκει το δείγμα. Η Εξίσωση 3.4 μπορεί πλέον να γραφεί και με την παρακάτω μορφή

$$p(C_k|X) = \frac{p(C_k)\prod_{i=1}^{n} p(x_i|C_k)}{p(X)} \tag{3.6}$$

Η πιθανότητα $p(X)$ στην Εξίσωση 3.6 είναι μια σταθερά για γνωστό διάνυσμα χαρακτηριστικών X. Επομένως, η πιθανότητα $p(C_k|X)$ το δείγμα X να ανήκει στην κατηγορία C_k είναι ανάλογη μόνο του αριθμητή.

$$p(C_k|X) \propto p(C_k)\prod_{i=1}^{n} p(x_i|C_k) \tag{3.7}$$

όπου $p(C_k)$ είναι η σχετική συχνότητα της κλάσης στο σύνολο εκπαίδευσης.

Από την Εξίσωση 3.7 μπορεί να προκύψει ταξινομητής που επιλέγει την κλάση με την μεγαλύτερη πιθανότητα για το συγκεκριμένο δείγμα όπως φαίνεται παρακάτω.

$$y = \arg\max_y p(y) \prod_{i=1}^{n} p(x_i|y) \tag{3.8}$$

Από τον τρόπο υπολογισμού της πιθανότητας $p(x_i|y)$ προκύπτουν διάφοροι τύποι μπεϋζιανών ταξινομητών *επιβλεπομένης μάθησης* (supervised learning) που θα παρουσιαστούν με λεπτομέρεια παρακάτω.

Είναι σημαντικό να τονιστεί ότι η πολύ απλή αλλά και ισχυρή προϋπόθεση, πως τα χαρακτηριστικά του διανύσματος είναι ανεξάρτητα μεταξύ τους, δεν ισχύει σε πολλά σύγχρονα προβλήματα στο πεδίο της μηχανικής μάθησης. Παρ' όλα αυτά, οι μπεϋζιανοί ταξινομητές συμπεριφέρονται αρκετά καλά σε αρκετές περιπτώσεις, λ.χ. κατηγοριοποίησης κειμένου. Πλεονέκτημα τους σε σχέση με άλλους ταξινομητές είναι ότι η εκπαίδευση τους απαιτεί μικρό αριθμό δειγμάτων και γίνεται πολύ γρηγορότερα.

3.2.1 Γκαουσιανός απλός μπεϋζιανός ταξινομητής

Όταν έχουμε δεδομένα με συνεχείς τιμές, συνήθως, μπορούμε να υποθέσουμε ότι αυτές οι συνεχείς τιμές ακολουθούν την γκαουσιανή κατανομή.

Ας υποθέσουμε ότι τα δεδομένα εκπαίδευσης για ένα πρόβλημα, περιέχουν ένα χαρακτηριστικό x με συνεχείς τιμές. Αρχικά, χωρίζουμε τα δεδομένα ανάλογα με την κατηγορία που ανήκουν και μετά υπολογίζουμε την μέση τιμή μ_c για το συγκεκριμένο χαρακτηριστικό και για δεδομένα που ανήκουν στην κλάση c. Παρομοίως, υπολογίζουμε και την διακύμανση σ_c^2 των τιμών του χαρακτηριστικού για την κάθε κλάση c. Τότε, μπορούμε να υπολογίσουμε την κατανομή πιθανότητας για μια τιμή $v, p(x = v|c)$ από τον παρακάτω τύπο.

$$p(x = v|c) = \frac{1}{\sqrt{s\pi\sigma_c^2}} \mathrm{e}^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}} \tag{3.9}$$

που είναι ο τύπος της κανονικής κατανομής με μέση τιμή μ_c και απόκλιση σ_c^2.

Στο Σχήμα 3.2 φαίνεται η γραφική παράσταση της κανονικής κατανομής ενός χαρακτηριστικού x με μέση τιμή μ_c και διακύμανση σ_c^2.

3.2.2 Απλός μπεϋζιανός ταξινομητής κατανομής Bernulli

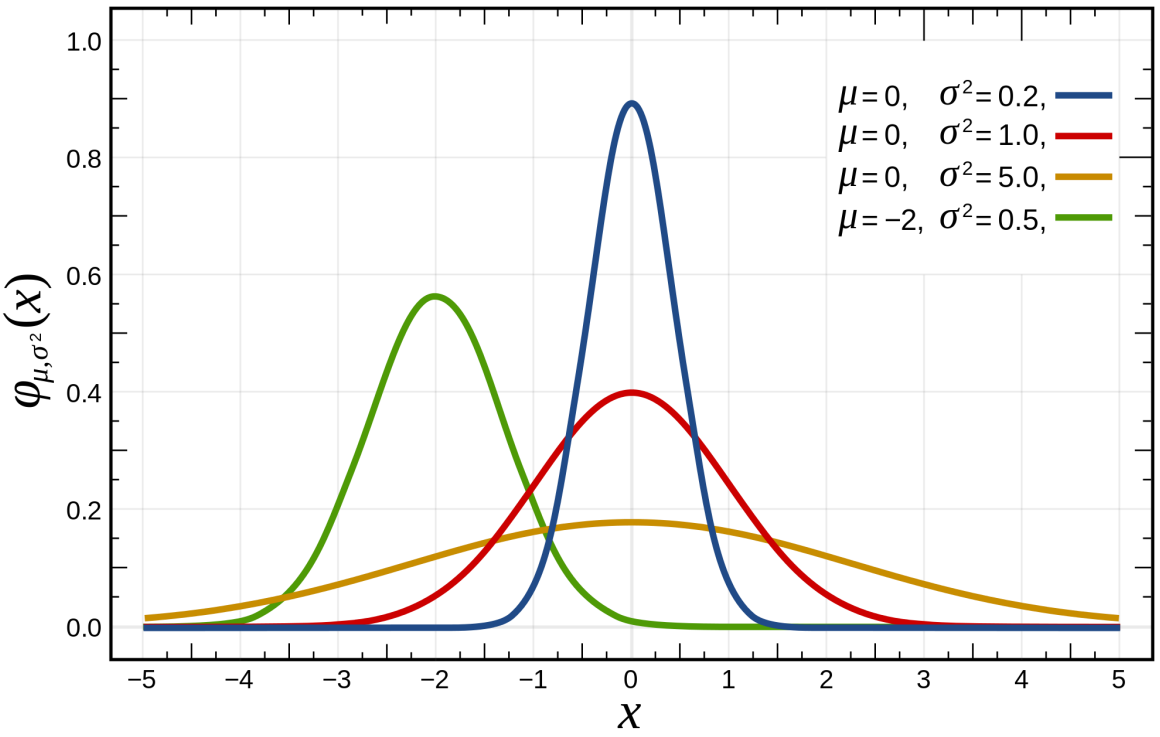
Σε ένα μοντέλο γεγονότων Bernulli, τα χαρακτηριστικά εισόδου είναι ανεξάρτητες λογικές, δυαδικές μεταβλητές. Αν υποθέσουμε ένα διάνυσμα εισόδου $X = (x_1, \ldots, x_n)$ με το χαρακτηριστικό x_i να δηλώνει την ύπαρξη του ή μη στο πρόβλημα (μόνο δυαδικές τιμές), τότε το δείγμα X έχει πιθανότητα $p(X|C_k)$ να ανήκει στην κλάση C_k και δίνεται από τον τύπο [McCa98]

$$p(X|C_k) = \prod_{i=1}^{n} p_{k_i}^{x_i} (1 - p_{k_i})^{1-x_i} \tag{3.10}$$

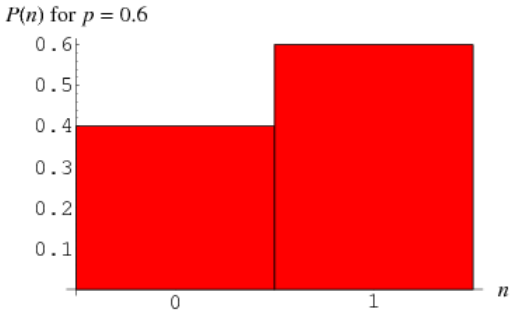
όπου p_{k_i} είναι η πιθανότητα το χαρακτηριστικό x_i να ανήκει στην κλάση C_k.

Η κατανομή Bernulli είναι μια διακριτή κατανομή με δυο δυνατά αποτελέσματα, την επιτυχία ($n = 1$) και την αποτυχία ($n = 0$). Η επιτυχία μπορεί να συμβεί με πιθανότητα p, ενώ η αποτυχία μπορεί να συμβεί με πιθανότητα $q \equiv 1 - p$, όπου $0 < p < 1$. Η συνάρτηση πυκνότητας πιθανότητας για ένα χαρακτηριστικό x_i που ακολουθεί την κατανομή Bernulli μπορεί να φανεί στην Εξίσωση 3.11 και το Σχήμα 3.3.

$$P(n) = \begin{cases} 1 - p, \text{αν } n = 0 \\ p, \text{αν } n = 1 \end{cases} \tag{3.11}$$



Σχήμα 3.2: Γραφική παράσταση κανονικής κατανομής χαρακτηριστικού x



Σχήμα 3.3: Γραφική παράσταση πυκνότητας πιθανότητας κατανομής Bernulli χαρακτηριστικού x

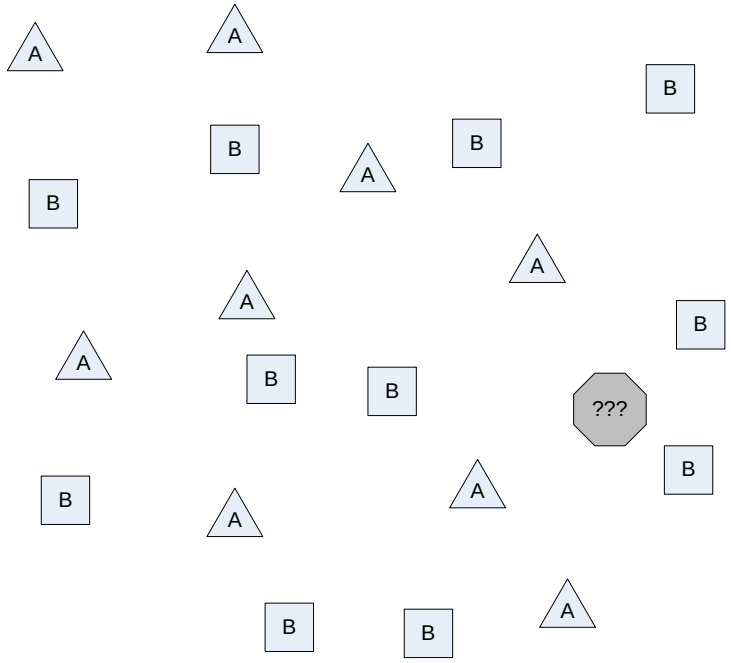
Ο απλός μπεϋζιανός ταξινομητής κατανομής Bernulli χρησιμοποιείται κυρίως για κατηγοριοποίηση κειμένου [McCa98].

3.3 Ταξινομητής k-πλησιέστερων γειτόνων

Ο ταξινομητής των *k-πλησιέστερων γειτόνων* (k-nearest neighbors classifier) εντάσσεται στην κατηγορία των μεθόδων επιβλεπομένης μάθησης. Η λειτουργία του είναι αρκετά απλή, μιας και κάθε νέο δείγμα τοποθετείται στην κλάση που ανήκουν τα k πλησιέστερα σε αυτό δείγματα, με την απόσταση μεταξύ των δειγμάτων να υπολογίζεται με κάποια από τις γνωστές μετρικές (λ.χ. Ευκλείδεια απόσταση). Ο αριθμός k αποτελεί παράμετρο του ταξινομητή.

Παρά την απλότητα του, ο ταξινομητής χρησιμοποιείται με επιτυχία σε μεγάλο αριθμό προβλημάτων ταξινόμησης, όπως προβλήματα αναγνώρισης γραμμάτων και αριθμών γραμμένα από ανθρώπινο χέρι. Είναι σημαντικό να τονιστεί ότι ο ταξινομητής k- πλησιέστερων γειτόνων μπορεί να έχει καλή απόδοση σε περιπτώσεις όπου το *σύνορο απόφασης* (decision boundary) μεταξύ των κλάσεων στον χώρο αναπαράστασης των δειγμάτων είναι δύσκολο να βρεθεί [Brem05] (Σχήμα 3.4).

Στο πρόβλημα του Σχήματος 3.4 υπάρχουν δυο κλάσεις Α και Β, αλλά είναι πολύ δύσκολο να



Σχήμα 3.4: Παράδειγμα ταξινόμησης όπου τα δείγματα είναι δύσκολο να διαχωριστούν σε ομάδες

βρεθεί διαχωριστική γραμμή που να τις χωρίζει σε δυο ευδιάκριτες ομάδες. Ο αλγόριθμος των k-πλησιέστερων γειτόνων μπορεί να βρει πολύ εύκολα σε ποια κλάση ανήκει το νέο δείγμα με σχήμα εξαγώνου.

3.3.1 Ο αλγόριθμος k-πλησιέστερων γειτόνων

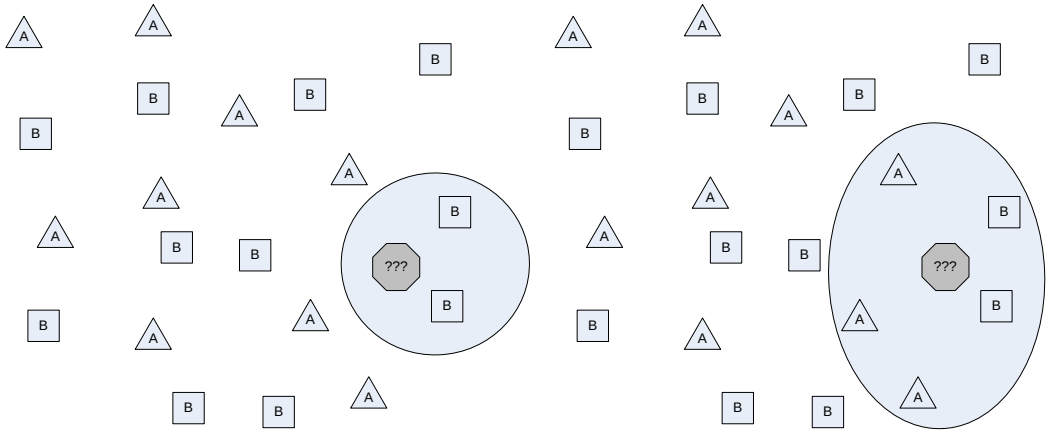
Ας υποθέσουμε ότι έχουμε πολλά δείγματα, τα οποία αποτελούν το σύνολο εκπαίδευσης του ταξινομητή. Τα δείγματα αυτά είναι διανύσματα χαρακτηριστικών και αναπαριστώνται στον πολυδιάστατο χώρο, όπως προαναφέραμε και στην εισαγωγή του Κεφαλαίου. Επίσης, κάθε δείγμα συνοδεύεται από την ετικέτα με την κλάση στην οποία ανήκει.

Στην φάση εκπαίδευσης του αλγορίθμου, δεν χρειάζεται να φτιαχτεί ένα μοντέλο προβλέψεων όπως γίνεται στον μπεϋζιανό ταξινομητή και σε πολλούς άλλους ταξινομητές. Ο αλγόριθμος, κατά την φάση της ταξινόμησης νέων δειγμάτων υπολογίζει τοπικά τους πλησιέστερους γειτόνους και αποφασίζει σε ποια κλάση ανήκει το νέο δείγμα. Ο συγκεκριμένος τρόπος μάθησης, ονομάζεται *οκνηρή μάθηση* (lazy learning), καθώς οι διάφοροι υπολογισμοί γίνονται τη στιγμή της ταξινόμησης του νέου δείγματος. Αυτό καθιστά τον αλγόριθμο k-πλησιέστερων γειτόνων σε έναν από τους απλούστερους αλγορίθμους μηχανικής μάθησης.

3.3.2 Παράμετροι του αλγορίθμου

Αριθμός γειτόνων Όπως αναφέρθηκε στην εισαγωγή, ο αλγόριθμος των k-πλησιέστερων γειτόνων, βρίσκει τους πιο κοντινούς γείτονες του νέου δείγματος και ανάλογα με την κλάση στην οποία ανήκουν, ταξινομεί το νέο δείγμα κατάλληλα. Μια πολύ σημαντική παράμετρος του αλγορίθμου είναι ο αριθμός k των γειτόνων που πρέπει να ελεγχθούν (Σχήμα 3.5).

Στο Σχήμα 3.5a ο αριθμός των γειτόνων που ελέγχονται έχει οριστεί στην τιμή $k = 2$ και εφόσον οι δυο πλησιέστεροι γείτονες ανήκουν στην κλάση B, τότε και το νέο δείγμα θα ανήκει στην κλάση B. Αντίθετα, στο Σχήμα 3.5b, ο αριθμός των ελέγξιμων γειτόνων αυξάνεται σε $k = 5$ και τότε επειδή τρεις γείτονες ανήκουν στην κλάση A έναντι δυο γειτόνων στην κλάση B, τελικά το νέο δείγμα τοποθετείται στην κλάση A.



(a) Αριθμός γειτόνων $k = 2$ **(b)** Αριθμός γειτόνων $k = 5$

Σχήμα 3.5: Φάση ταξινόμησης του αλγορίθμου k-πλησιέστερων γειτόνων.

Υπολογισμός πλησιέστερων γειτόνων Γνωρίζουμε ότι τα δείγματα είναι διανύσματα στον χώρο. Ο υπολογισμός της απόστασης μεταξύ δυο δειγμάτων είναι ουσιαστικά ο υπολογισμός της απόστασης μεταξύ δυο διανυσμάτων. Ο υπολογισμός αυτός μπορεί να γίνει χρησιμοποιώντας την Ευκλείδεια απόσταση ή την απόσταση Manhattan.

Αν υποτεθούν δυο διανύσματα $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ και $\mathbf{q} = (q_1, q_2, \ldots, q_n)$, ο υπολογισμός της Ευκλείδειας απόστασης μεταξύ τους μπορεί να φανεί στην Εξίσωση 3.12, ενώ ο υπολογισμός της απόστασης Manhattan στην Εξίσωση 3.13.

$$\mathcal{D}_{eucl}(\mathbf{p}, \mathbf{q}) \equiv \|\mathbf{p} - \mathbf{q}\|_2 = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{3.12}$$

$$\mathcal{D}_{man}(\mathbf{p}, \mathbf{q}) \equiv \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^{n} |p_i - q_i| \tag{3.13}$$

Για την εύρεση των k πλησιέστερων γειτόνων μπορεί να πραγματοποιηθεί *εξαντλητική αναζήτηση* (exhaustive search) μεταξύ όλων των ζευγαριών των δειγμάτων. Η χρονική πολυπλοκότητα αυτής της μεθόδου είναι $\mathcal{O}(dn^2)$ για n αριθμό δειγμάτων και d διάσταση του διανύσματος χαρακτηριστικών. Για μεγάλο αριθμό δειγμάτων n η εξαντλητική αναζήτηση είναι απαγορευτική.

Η χρήση δεντρικών δομών βοηθά στην απόδοση έναντι της εξαντλητικής αναζήτησης. Η βασική ιδέα είναι να μην ελεγχθούν όλα τα δυνατά ζευγάρια δειγμάτων, αλλά αν είναι γνωστό ότι το δείγμα Α είναι πολύ μακριά από το δείγμα Β και είναι επίσης γνωστό ότι το δείγμα Β είναι πολύ κοντά στο δείγμα Γ, τότε γνωρίζουμε επίσης ότι το δείγμα Α είναι πολύ μακριά και από το δείγμα Γ. Με αυτόν τον τρόπο το υπολογιστικό κόστος εύρεσης των πλησιέστερων γειτόνων μπορεί να ελαττωθεί σε $\mathcal{O}(dn \log n)$ ή και παρακάτω.

Βάρη σημαντικότητας των γειτόνων Μέχρι τώρα θεωρήθηκαν ομοιόμορφα βάρη για τους γείτονες ενός νέου δείγματος. Αυτό σημαίνει πως αν ελεγχθούν οι k πλησιέστεροι γείτονες, η κλάση του νέου δείγματος, είναι η κλάση που ανήκουν οι περισσότεροι από τους γειτόνους του και η συνεισφορά κάθε γείτονα έχει βαρύτητα 1.

Πολλές φορές είναι χρήσιμο να τίθενται βάρη στη συνεισφορά κάθε γείτονα. Για παράδειγμα, η σημαντικότητα της συνεισφοράς ενός γείτονα μπορεί να είναι αντιστρόφως ανάλογη με την απόσταση του γείτονα από το εξεταζόμενο προς ταξινόμηση νέο δείγμα. Με αυτόν τον τρόπο, γείτονες που βρίσκονται πιο κοντά στο νέο δείγμα επηρεάζουν περισσότερο την απόφαση της κλάσης που ανήκει το δείγμα. Τα βάρη στην συνεισφορά των γειτόνων μπορεί να έχουν μεγάλη σημασία σε συγκεκριμένες περιπτώσεις.

Ας υποτεθεί η περίπτωση όπου μια κλάση υπερεκπροσωπείται έναντι των άλλων. Ένα νέο δείγμα, είναι πολύ πιθανόν να έχει ως γείτονες πάρα πολλά δείγματα της κλάσης που υπερεκπροσωπείται και άρα αν τα βάρη της συνεισφοράς των γειτόνων είναι ομοιόμορφα, το νέο δείγμα θα ανήκει και αυτό στην υπερισχύουσα κλάση. Αντίθετα, αν η σημαντικότητα της συνεισφοράς κάθε γείτονα είναι αντιστρόφως ανάλογη του ποσοστού ύπαρξης της κλάσης, τότε γείτονες που ανήκουν σε ανίσχυρες κλάσεις θα έχουν ισοδύναμη συνεισφορά με τους γείτονες που ανήκουν στην υπερισχύουσα κλάση [Coom82].

3.4 Δέντρα αποφάσεων

Τα *δέντρα αποφάσεων* (decision trees) είναι μια μέθοδος επιβλεπομένης μάθησης που χρησιμοποιείται κυρίως στην στατιστική, στην εξόρυξη δεδομένων και στην μηχανική μάθηση [Roka08]. Σκοπός ενός δέντρου αποφάσεων είναι να δημιουργήσει ένα μοντέλο προβλέψεων το οποίο αντιστοιχίζει παρατηρήσεις για τα χαρακτηριστικά ενός δείγματος σε μια κλάση για το συγκεκριμένο δείγμα. Μοντέλα δέντρων που οι κλάσεις του προβλήματος έχουν συγκεκριμένες τιμές από ένα σύνολο, ονομάζονται δέντρα ταξινόμησης και κατηγοριοποίησης. Σε αυτά τα δέντρα τα τελικά φύλλα αντιστοιχούν στις ετικέτες των κλάσεων του προβλήματος και τα διάφορα μονοπάτια πάνω στο δέντρο αντιστοιχούν σε λογικούς συνδυασμούς που έχουν γίνει για τα χαρακτηριστικά του δείγματος.

3.4.1 Κατασκευή δέντρου αποφάσεων

Ας υποθέσουμε ότι τα δεδομένα εκπαίδευσης έχουν την μορφή διανύσματος

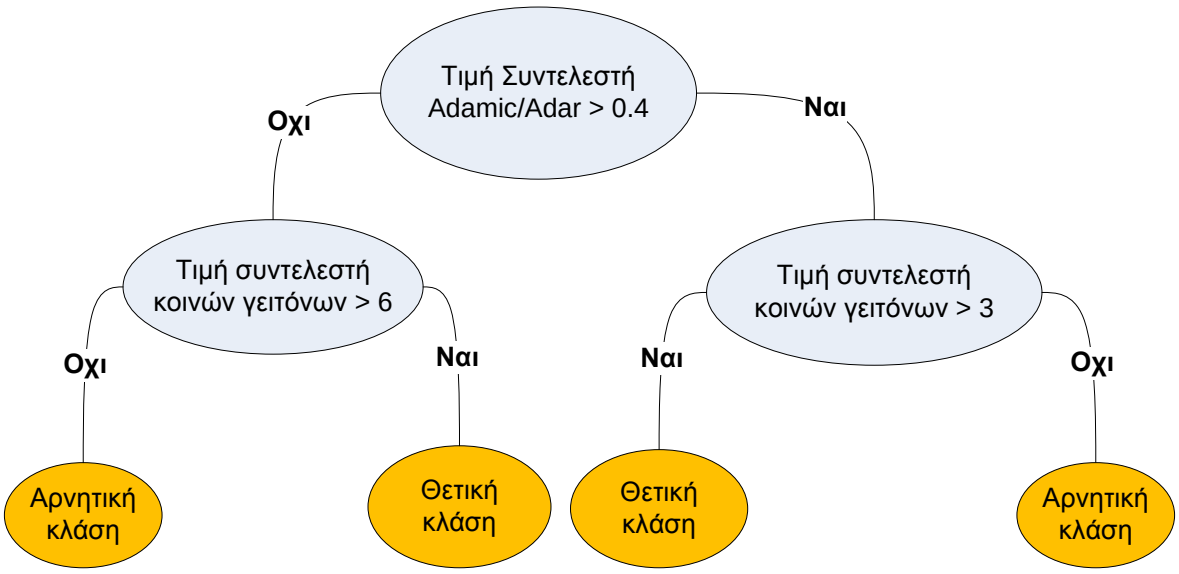
$$(X, Y) = (x_1, x_2, \ldots, x_k, Y) \tag{3.14}$$

Η μεταβλητή Y είναι η ετικέτα που υποδηλώνει σε ποια κλάση του προβλήματος ανήκει το δείγμα. Οι μεταβλητές $x_1, x_2, \ldots, x_k$ είναι τα διάφορα χαρακτηριστικά του δείγματος. Ο σκοπός της κατασκευής του δέντρου είναι η δημιουργία ενός μοντέλου που θα προβλέπει την κλάση Y ενός δείγματος με βάση όλα τα χαρακτηριστικά εισόδου $x_1, x_2, \ldots, x_k$. Ένα παράδειγμα δέντρου αποφάσεων φαίνεται στο Σχήμα 3.6.

Το παράδειγμα αναφέρεται σε ένα μέσο κοινωνικής δικτύωσης για το οποίο έχει κατασκευαστεί ένα πολύ απλό δέντρο αποφάσεων. Τα δείγματα του παραδείγματος, αναπαριστούνται από ένα διάνυσμα που περιέχει μόνο δυο χαρακτηριστικά, τον συντελεστή Adamic-Adar και τον συντελεστή των κοινών γειτόνων. Με πορτοκαλί χρώμα επισημαίνονται τα φύλλα του δέντρου, τα οποία αποτελούν και τις κλάσεις του προβλήματος. Οι εσωτερικοί κόμβοι του δέντρου, δηλαδή οι κόμβοι που δεν είναι φύλλα, αναφέρονται στα χαρακτηριστικά των δειγμάτων και οδηγούν σε διαφορετικά λογικά μονοπάτια ανάλογα με την τιμή του αντίστοιχου χαρακτηριστικού. Κάθε νέο δείγμα που πρέπει να ταξινομηθεί, θα περάσει από ένα μονοπάτι του δέντρου αρχίζοντας από τον αρχικό κόμβο / ρίζα και καταλήγοντας σε έναν τελικό κόμβο / φύλλο του δέντρου. Το φύλλο που θα καταλήξει το δείγμα, δηλώνει και την κλάση στην οποία ανήκει.

Με άλλα λόγια, κάθε χαρακτηριστικό των δειγμάτων εισόδου έχει ένα πεδίο τιμών και αναπαριστάται με έναν εσωτερικό κόμβο στο δέντρο αποφάσεων. Από κάθε εσωτερικό κόμβο, μπορούμε να οδηγηθούμε σε δυο ή περισσότερους κόμβους / παιδιά που ο κάθε κόμβος υποδηλώνει ένα υποσύνολο του πεδίου τιμών του αρχικού κόμβου / πατέρα. Έτσι, ξεκινώντας από έναν κόμβο ρίζα και χωρίζοντας σε κάθε επίπεδο του δέντρου το πεδίο τιμών σε υποσύνολα αναδρομικά, κατασκευάζουμε το δέντρο αποφάσεων μέχρι το σημείο όπου θα φτάσουμε στα φύλλα του δέντρου και το πεδίο τιμών θα έχει μόνο μια τιμή, μια από τις κλάσεις του προβλήματος [Quin86].

Υπάρχουν πολλοί τύποι δέντρων αποφάσεων. Ο πιο γνωστός τύπος σε προβλήματα κατηγοριοποίησης είναι τα δέντρα CART (Classification and Regression Trees) [Brei84]. Ο αλγόριθμος CART δημιουργεί δυαδικά δέντρα αποφάσεων χρησιμοποιώντας μια τιμή κατωφλίου για κάθε χαρακτηριστικό, τέτοια ώστε να μεγιστοποιείται η *πληροφοριακή απολαβή* (information gain) σε κάθε κόμβο του δέντρου.



Σχήμα 3.6: Παράδειγμα δέντρου αποφάσεων

3.4.2 Αλγόριθμοι κατασκευής δέντρων αποφάσεων

Οι αλγόριθμοι κατασκευής δέντρων αποφάσεων συνήθως λειτουργούν με *από πάνω προς τα κάτω* (top-down) μεθοδολογία [Roka05]. Σε κάθε βήμα επιλέγουν μια μεταβλητή / χαρακτηριστικό η οποία χωρίζει με τον καλύτερο τρόπο το σύνολο των δειγμάτων εισόδου. Κάθε αλγόριθμος χρησιμοποιεί ξεχωριστές μετρικές για να υπολογίσει ποιος είναι ο καλύτερος τρόπος διαχωρισμού των δειγμάτων.

Δείκτης Gini (Gini index) Ένας τρόπος εύρεσης του καλύτερου διαχωρισμού του συνόλου δειγμάτων για ένα χαρακτηριστικό (εσωτερικός κόμβος του δέντρου αποφάσεων) είναι ο δείκτης Gini. Ο δείκτης Gini μπορεί να υπολογιστεί προσθέτοντας την πιθανότητα f_i κάθε δείγματος να επιλεγεί, επί την πιθανότητα $1 - f_i$ αυτό το δείγμα να μη κατηγοριοποιηθεί σωστά. Ο δείκτης μηδενίζεται όταν όλα τα δείγματα που εξετάζονται σε έναν εσωτερικό κόμβο του δέντρου κατηγοριοποιηθούν σε μια μόνο κατηγορία.

Για τον υπολογισμό του δείκτη Gini σε ένα σύνολο m δειγμάτων, υποθέτουμε $i \in \{1, 2, \ldots, m\}$ και θέτουμε το ποσοστό των δειγμάτων που έχουν τιμή μέσα στο σύνολο. Ο δείκτης προκύπτει από την παρακάτω Εξίσωση

$$I_G(f) = \sum_{i=1}^{m} f_i(1 - f_i) = \sum_{i=1}^{m}(f_i - f_i^2) = \sum_{i=1}^{m} f_i - \sum_{i=1}^{m} f_i^2 = 1 - \sum_{i=1}^{m} f_i^2 = \sum_{i=1}^{m} f_i f_k \tag{3.15}$$

Πληροφοριακή απολαβή (Information gain) Ένας άλλος τρόπος διαχωρισμού του συνόλου δειγμάτων είναι ο δείκτης πληροφοριακής απολαβής. Η ιδέα του υπολογισμού αυτού του δείκτη προκύπτει από την έννοια της εντροπίας στην θεωρία της πληροφορίας. Η εντροπία υπολογίζεται όπως παρακάτω (Εξίσωση 3.16)

$$H(f) = -\sum_{i=1}^{m} f_i \log_2 f_i \tag{3.16}$$

Κατόπιν, η πληροφοριακή απολαβή υπολογίζεται ως εξής (Εξίσωση 3.17)

$$\begin{aligned} \text{Κέρδος Πληροφορίας} &= \text{Εντροπία πατέρα} - \text{Άθροισμα εντροπίας παιδιών} \\ I_G(T, \alpha) &= H(T) - H(T|\alpha) \end{aligned} \tag{3.17}$$

3.4.3 Πλεονεκτήματα και μειονεκτήματα της χρήσης δένδρων αποφάσεων

Στην παρακάτω λίστα συνοψίζονται ορισμένα από τα πλεονεκτήματα της χρήσης δένδρων αποφάσεων έναντι άλλων τεχνικών ταξινόμησης.

1. Τα δέντρα αποφάσεων είναι απλά στην κατανόηση τους. Χρησιμοποιούν κυρίως κανόνες λογικής που είναι απλοί στο να γίνουν κατανοητοί και από τον άνθρωπο. Επίσης, είναι πολύ εύκολο να αναπαρασταθούν, όπως φαίνεται και στο παράδειγμα του Σχήματος 3.6.

2. Δεν απαιτείται προεπεξεργασία των δεδομένων εισόδου, όπως λ.χ. κανονικοποίηση.

3. Από την στιγμή που θα φτιαχτεί το δέντρο αποφάσεων, ο χρόνος προσπέλασής του είναι λογαριθμικός ως προς την είσοδο του ταξινομητή.

4. Τα δέντρα αποφάσεων μπορούν να χειριστούν αριθμητικά αλλά και *κατηγορικά* (categoral) δεδομένα.

5. Τα δέντρα αποφάσεων θεωρούνται μοντέλα άσπρου κουτιού, αφού κάθε μέρος του μοντέλου μπορεί να παρατηρηθεί και να αναλυθεί με λογικούς κανόνες.

6. Λειτουργούν καλά και σε προβλήματα με πολλά δεδομένα.

Παρακάτω παρουσιάζονται ορισμένα μειονεκτήματα της χρήσης δέντρων αποφάσεων.

1. Οι αλγόριθμοι κατασκευής των δέντρων αποφάσεων μπορεί να δημιουργήσουν πολύ περίπλοκα δέντρα που κατηγοριοποιούν σωστά τα δεδομένα εκπαίδευσης αλλά η απόδοση τους μειώνεται δραματικά κατά την φάση του ελέγχου. Το πρόβλημα αυτό ονομάζεται *υπερπροσαρμογή* (overfitting) και μπορεί να αντιμετωπιστεί με *κλάδεμα* (pruning) του δέντρου ή με επιβολή μέγιστου βάθους στο δέντρο [Hoth06].

2. Μικρές αλλαγές στα δεδομένα εκπαίδευσης μπορεί να προκαλέσουν πολλά προβλήματα σε ένα ήδη εκπαιδευμένο δέντρο αποφάσεων. Ένα νέο μοντέλο είναι αναγκαίο να δημιουργηθεί για να ανταποκρίνεται καλύτερα στα αλλαγμένα δεδομένα.

3. Το πρόβλημα της εύρεσης του βέλτιστου δέντρου αποφάσεων είναι ένα γνωστό NP-πλήρες πρόβλημα [Hyaf76]. Αυτό σημαίνει ότι χρησιμοποιούνται ευριστικοί αλγόριθμοι για τη λήψη τοπικά βέλτιστων αποφάσεων σε κάθε βήμα κατασκευής του δέντρου. Το τελικό δέντρο που κατασκευάζεται δεν είναι το ολικά βέλτιστο.

4. Υπάρχουν προβλήματα που είναι πολύ δύσκολο να αναλυθούν από ένα δέντρο αποφάσεων. Για παράδειγμα το πρόβλημα της δυαδική ισοτιμίας (πρόβλημα XOR) οδηγεί στην δημιουργία ενός πολύ περίπλοκου δέντρου, ενώ υπάρχουν άλλοι αλγόριθμοι που μπορούν να το λύσουν πολύ εύκολα.

5. Αν τα δεδομένα εισόδου είναι μη-ισορροπημένα, δηλαδή μια κλάση υπερτερεί σε αριθμό δειγμάτων έναντι των άλλων κλάσεων, τότε το δέντρο που θα δημιουργηθεί θα δίνει λανθασμένα μεγαλύτερη βαρύτητα στην υπερέχουσα κλάση.

3.5 Μηχανές Διανυσμάτων Υποστήριξης

Στο πεδίο της μηχανικής μάθησης, οι *μηχανές διανυσμάτων υποστήριξης* (support vector machines) είναι μοντέλα επιβλεπομένης μάθησης που χρησιμοποιούνται σε προβλήματα κατηγοριοποίησης. Στα μοντέλα αυτά, τα δείγματα αναπαριστώνται με σημεία στον χώρο και χωρίζονται από ένα υπερεπίπεδο σε δυο κατηγορίες. Η διαδικασία της ταξινόμησης νέων δειγμάτων απαιτεί την αναπαράσταση των δειγμάτων αυτών στον χώρο και την παρατήρηση σε ποια μεριά του υπερεπιπέδου βρίσκονται, η με άλλα λόγια σε ποια κατηγορία εντάσσονται.

3.5.1 Διαδικασία μάθησης

Η διαδικασία μάθησης ενός μοντέλου μηχανών διανυσμάτων υποστήριξης είναι ουσιαστικά η επίλυση ενός προβλήματος βελτιστοποίησης. Τα δείγματα που χρησιμοποιούνται για την εκπαίδευση του μοντέλου και για την μέτρηση της απόδοσης του, είναι διανύσματα διάστασης d και μπορούν να αναπαρασταθούν με σημεία στο χώρο. Κάθε δείγμα ανήκει σε μια κατηγορία του προβλήματος, την θετική ή την αρνητική κατηγορία. Έτσι, το σύνολο εκπαίδευσης αποτελείται από ζευγάρια (x_i, y_i), όπου $x_i \in R^d$ είναι το διάνυσμα του i-οστού δείγματος και $y_i \in +1, -1$ είναι η ετικέτα της κατηγορίας που ανήκει το i-οστό δείγμα [Cort95].

Το μοντέλο προσπαθεί να διαχωρίσει τα θετικά δείγματα από τα αρνητικά χρησιμοποιώντας ένα υπερεπίπεδο της μορφής $w^\top x = b$. Το w είναι ένα διάνυσμα κάθετο στο υπερεπίπεδο που δείχνει ποια είναι η διεύθυνση του επιπέδου. Η σταθερά b καθορίζει την θέση του επιπέδου στον χώρο. Τα ιδανικά w και b βρίσκονται επιλύοντας το πρόβλημα βελτιστοποίησης των Εξισώσεων 3.18-3.19.

$$\arg\min_{w} \left[\frac{1}{2}||w||^2 \quad + \quad C\sum_i \xi_i \right] \tag{3.18}$$

$$\begin{aligned} \text{subject to} \quad y_i(w^\top x - b) &\geq 1 - \xi_i, \\ \xi_i &\geq 0 \end{aligned} \tag{3.19}$$

Το κενό μεταξύ των επιπέδων $w^\top x = b \pm 1$ ονομάζεται διάκενο και εκφράζεται συναρτήσει του w ως $\frac{2}{||w||}$. Η συνάρτηση κόστους της Εξίσωσης 3.18 ελαχιστοποιείται, δηλαδή ο όρος $||w||^2$ επίσης ελαχιστοποιείται, ή αλλιώς μεγιστοποιείται το διάκενο και παράλληλα ελαχιστοποιείται το σφάλμα κατηγοριοποίησης, που είναι εκφρασμένο με τον όρο $\sum_i \xi_i$. Η σταθερά C παίρνει τιμή πριν την εκπαίδευση και υποδηλώνει την σχετική σημαντικότητα των δυο παραπάνω όρων. Ιδανικά, θέλουμε όλα τα δείγματα εκπαίδευσης να ταξινομούνται στην σωστή κατηγορία και μακριά από το διάκενο, έτσι ώστε ο όρος $\sum_i \xi_i$ να παραμένει μικρός και το διάκενο να είναι μεγάλο. Τα Σχήμα 3.7 βοηθά στην καλύτερη κατανόηση της διαδικασίας μάθησης των μηχανών διανυσμάτων υποστήριξης.

Στο Σχήμα 3.7a παρατηρούνται τρία πιθανά υπερεπίπεδα που δημιουργούνται κατά την διαδικασία εκπαίδευσης ενός μοντέλου μηχανών διανυσμάτων υποστήριξης. Οι δυο κλάσεις του προβλήματος φαίνονται στον δισδιάστατο χώρο με διαφορετικά χρώματα. Το υπερεπίπεδο H_1 αποτυγχάνει να ταξινομήσει σωστά τα δείγματα, ενώ το υπερεπίπεδο H_2 κάνει σωστή ταξινόμηση των δειγμάτων αλλά χωρίς να μεγιστοποιήσει το διάκενο γύρω από το υπερεπίπεδο. Το ιδανικό επίπεδο για το συγκεκριμένο πρόβλημα είναι το επίπεδο H_3. Παρατηρούμε ότι ταξινομεί σωστά όλα τα δείγματα και δημιουργεί το μέγιστο δυνατό διάκενο μεταξύ των κλάσεων.

Στο Σχήμα 3.7b φαίνεται πως βρίσκεται το μέγιστο διάκενο μεταξύ των δυο κλάσεων, αλλά και η σημαντικότητα της σταθεράς b στην διαδικασία μάθησης.

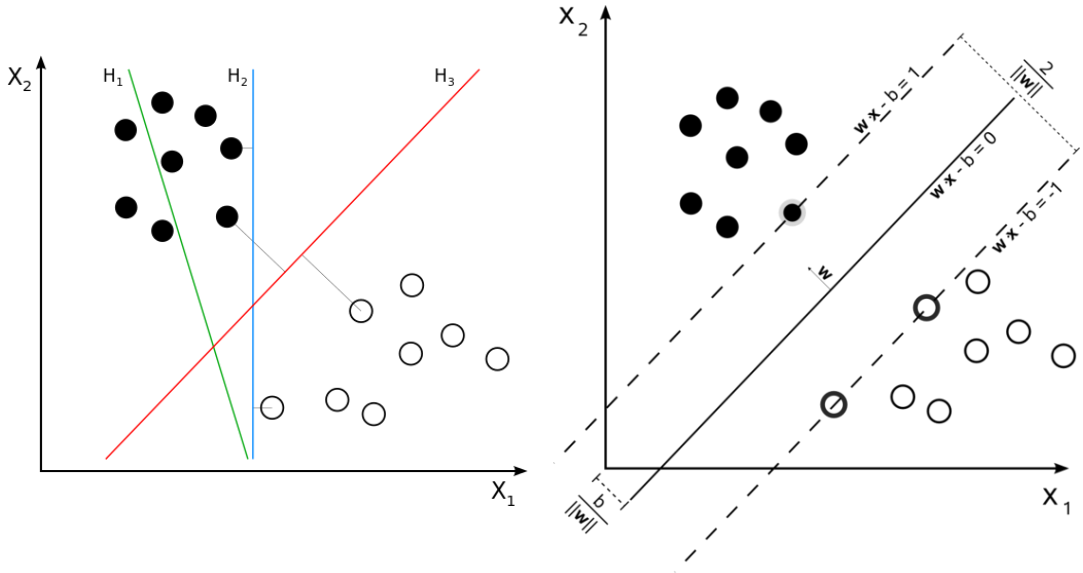
Αφού υπολογιστούν τα βέλτιστα w και b, ο ταξινομητής χρησιμοποιεί το παρακάτω κριτήριο για να ταξινομήσει νέα δείγματα.

$$\text{πρόβλεψη}(x) = \text{πρόσημο}(w^\top x - b) \tag{3.20}$$

Αν ένα διάνυσμα x έχει θετικό πρόσημο στην αποτίμηση της παραπάνω εξίσωσης, τότε ανήκει στην θετική κατηγορία. Η τιμή της σταθεράς b παίζει πολύ σημαντικό ρόλο.

3.5.2 Μη γραμμική κατηγοριοποίηση

Όπως ειπώθηκε πριν, τα δείγματα εκπαίδευσης έχουν μορφή διανυσμάτων διάστασης d και μπορούν να αναπαρασταθούν στον d-διάστατο χώρο των χαρακτηριστικών. Πολλές φορές δεν υπάρχει γραμμικός ταξινομητής που να ταξινομεί τα δείγματα, ή με άλλα λόγια δεν υπάρχει υπερεπίπεδο στον d-διάστατο χώρο που να διαχωρίζει σαφώς τις κατηγορίες στις οποίες εντάσσονται τα δείγματα.



(a) Διάφορα υπερεπίπεδα

(b) Διαδικασία εύρεσης του βέλτιστου υπερεπίπεδου

Σχήμα 3.7: Διαδικασία μάθησης Μηχανών Διανυσμάτων Υποστήριξης

Ωστόσο, οι [Bose92] πρότειναν μια λύση για την δημιουργία μη γραμμικών ταξινομητών με χρήση των μηχανών διανυσμάτων υποστήριξης. Η ιδέα είναι να αναπαρασταθούν τα δεδομένα σε χώρο μεγαλύτερης διάστασης χρησιμοποιώντας συναρτήσεις πυρήνα. Στο νέο χώρο υπάρχει υπερεπίπεδο που χωρίζει τα δεδομένα σε κλάσεις, ασχέτως αν το επίπεδο αυτό ήταν αδύνατο να βρεθεί στον αρχικό χώρο χαρακτηριστικών του προβλήματος.

3.5.3 Το πρόβλημα της ανισορροπίας των κλάσεων

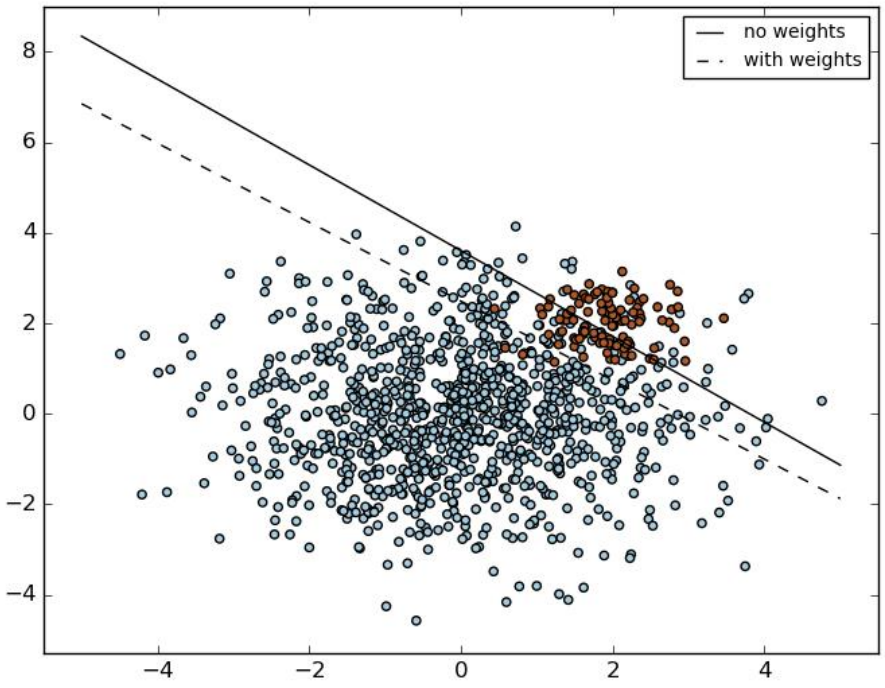
Σε πολλά προβλήματα τα δείγματα που ανήκουν στην θετική κατηγορία είναι πολύ λίγα σε σχέση με τα αρνητικά δείγματα. Αυτό ισχύει και στο πρόβλημα της πρόβλεψης ακμών στα μέσα κοινωνικής δικτύωσης που πραγματεύεται η παρούσα διπλωματική εργασία. Για παράδειγμα, στο Facebook οι υπαρκτές ακμές του δικτύου αποτελούν λιγότερο από το 10% όλων των πιθανών ακμών του δικτύου. Αυτό σημαίνει ότι η θετική κλάση έχει λιγότερα από το 10% των δειγμάτων του συνόλου εκπαίδευσης, ενώ η αρνητική κλάση, οι μη υπαρκτές ακμές του δικτύου δηλαδή, υπερισχύουν με ποσοστό άνω του 90%. Όπως είναι λογικό, η εκπαίδευση ενός συστήματος διανυσμάτων υποστήριξης με αυτά τα δεδομένα θα είναι μεροληπτική προς την αρνητική κατηγορία.

Οι [Bran03] προτείνουν την επιβολή βαρών σε κάθε κλάση, έτσι ώστε τα θετικά δείγματα να έχουν μεγαλύτερη βαρύτητα. Με αυτόν τον τρόπο επηρεάζεται έμμεσα ο υπολογισμός των παραμέτρων w και b. Ο νέος τύπος της συνάρτησης κόστους δίνεται παρακάτω.

$$f(w,b) = \frac{1}{2}||w||^2 + jC \sum_{i,y_i=1} \xi_i + C \sum_{i,y_i=-1} \xi_i \tag{3.21}$$

Η σταθερά j αυξάνει το κόστος της λανθασμένης κατηγοριοποίησης ενός θετικού δείγματος. Το Σχήμα 3.8 βοηθά στην κατανόηση της σημαντικότητας της νέας συνάρτησης κόστους (Εξίσωση 3.21)

Παρατηρούμε πως αν δεν χρησιμοποιηθούν βάρη κλάσεων πολλά από τα θετικά δείγματα κατηγοριοποιούνται σε λάθος κλάση, καθώς τα αρνητικά δείγματα είναι πολύ περισσότερα. Αντίθετα, σε προβλήματα που μας ενδιαφέρει κυρίως η σωστή κατηγοριοποίηση των θετικών δειγμάτων, όπως στο πρόβλημα που εξετάζεται στην παρούσα εργασία, το υπερεπίπεδο που υπολογίζεται με την χρήση βαρών στις κλάσεις, είναι σωστό.



Σχήμα 3.8: Χρήση βαρών κλάσεων κατά την φάση εκπαίδευσης συστήματος διανυσμάτων υποστήριξης

3.6 Στοχαστική Κάθοδος Κλίσης

Η *στοχαστική κάθοδος κλίσης* (stochastic gradient descent) είναι μια ακόμα μέθοδος επιβλεπομένης μάθησης. Η διαδικασία της εκπαίδευσης ταξινομητών με αυτή τη μέθοδο, ισοδυναμεί με την επίλυση ενός προβλήματος βελτιστοποίησης, όπως ακριβώς και στις μηχανές διανυσμάτων υποστήριξης. Όπως θα φανεί παρακάτω, η μέθοδος της στοχαστικής καθόδου κλίσης είναι πολύ αποτελεσματική σε προβλήματα με πολλά δεδομένα και πολλά χαρακτηριστικά [Bous08].

3.6.1 Διαδικασία μάθησης

Όπως και στις μηχανές διανυσμάτων υποστήριξης, ο βασικός στόχος είναι ο εντοπισμός ενός υπερεπιπέδου που χωρίζει τα δεδομένα σε κατηγορίες στον χώρο των χαρακτηριστικών.

Ας υποθέσουμε ότι έχουμε ένα σύνολο με δεδομένα εκπαίδευσης της μορφής $(x_1, y_1), \ldots, (x_n, y_n)$, όπου $x_i \in R^n$ και $y_i \in \{-1, 1\}$. Προσπαθούμε να βρούμε μια συνάρτηση της μορφής

$$f(x) = w^\top x + b \tag{3.22}$$

με παραμέτρους $w \in R^m$ και $b \in R$, τέτοια ώστε να μπορεί να προβλέπει σωστά την κατηγορία νέων δειγμάτων. Για δοσμένα w και b, αν ένα δείγμα x_i δώσει αρνητικό πρόσημο στην Εξίσωση 3.22, τότε ανήκει στην αρνητική κατηγορία, αλλιώς στην θετική [Zhan04b].

Ένας τρόπος εύρεσης των παραμέτρων του μοντέλου είναι η ελαχιστοποίηση του κανονικοποιημένου σφάλματος εκπαίδευσης (Εξίσωση 3.23).

$$E(w, b) = \frac{1}{n} \sum_{i=1}^{n} L\left(y_i, f(x_i)\right) + \alpha R(w) \tag{3.23}$$

Η συνάρτηση σφάλματος L υπολογίζει το σφάλμα μεταξύ της πρόβλεψης $f(x_i)$ και της πραγματικής κατηγορίας y_i για κάθε δεδομένο εκπαίδευσης x_i. Το R είναι όρος κανονικοποίησης που εμποδίζει την μεγάλη πολυπλοκότητα του μοντέλου και το φαινόμενο της υπερπροσαρμογής και τέλος η παράμετρος α είναι πάντα θετική.

Διαφορετικές συναρτήσεις σφάλματος L δίνουν διαφορετικούς ταξινομητές, ενώ ο όρος κανονικοποίησης μπορεί να υπολογιστεί και αυτός με διάφορους τρόπους, όπως φαίνεται παρακάτω.

$$\text{L2-norm} \quad R(w) = \frac{1}{2} \sum_{i=1}^{n} w_i^2 \tag{3.24}$$

$$\text{L1-norm} \quad R(w) = \sum_{i=1}^{n} |w_i| \tag{3.25}$$

Η μέθοδος της στοχαστικής καθόδου κλίσης προσπαθεί να ελαχιστοποιήσει τη συνάρτηση σφάλματος (Εξίσωση 3.23), υπολογίζοντας σε κάθε βήμα του αλγορίθμου την παράγωγο της $E(w, b)$ μόνο για ένα δείγμα εκπαίδευσης τη φορά. Ο αλγόριθμος εξετάζει όλα τα δείγματα που βρίσκονται στο σύνολο εκπαίδευσης και για κάθε δείγμα που εξετάζει, ενημερώνει τις παραμέτρους του μοντέλου σύμφωνα με τον παρακάτω κανόνα.

$$w \leftarrow w - \eta \left(\alpha \frac{\partial R(w)}{\partial w} + \frac{\partial L(y_i, f(x_i))}{\partial w} \right) \tag{3.26}$$

όπου η είναι ο ρυθμός μάθησης που μπορεί να είναι σταθερός ή να μειώνεται σταδιακά καθώς ο αλγόριθμος προχωρά. Η παράμετρος b υπολογίζεται με τον ίδιο τρόπο αλλά χωρίς τον όρο κανονικοποίησης.

Το γεγονός ότι η ενημέρωση των παραμέτρων του μοντέλου γίνεται μόνο για ένα δείγμα σε κάθε βήμα του αλγορίθμου, καθιστά την μέθοδο της στοχαστικής καθόδου κλίσης πολύ αποδοτική ακόμα και σε μεγάλων διαστάσεων προβλήματα [Bous08]. Ο συνολικός αριθμός των επαναλήψεων που θα εξεταστούν τα δείγματα του συνόλου εκπαίδευσης (epochs), καθώς και ο *ρυθμός μάθησης* (learning rate) είναι παράμετροι που θα πρέπει να καθοριστούν πριν την έναρξη του αλγορίθμου.

Κεφάλαιο 4

Πειραματική διαδικασία

Σε αυτό το Κεφάλαιο θα περιγραφεί ο τρόπος με τον οποίο προσεγγίστηκε το πρόβλημα της πρόβλεψης ακμών στα μέσα κοινωνικής δικτύωσης σε αυτήν την εργασία. Αρχικά, θα δοθούν πληροφορίες για τα σύνολα δεδομένων που χρησιμοποιήθηκαν. Στην συνέχεια, θα περιγραφούν τα πειραματικά πρωτόκολλα, οι τεχνικές που υλοποιήθηκαν καθώς και ο τρόπος μέτρησης της απόδοσης της κάθε τεχνικής.

Είναι σημαντικό να τονιστεί ότι το πειραματικό πρωτόκολλο των τεχνικών που βασίζονται στη γραφοθεωρία (Κεφάλαιο 2) είναι διαφορετικό από αυτό των ευφυών τεχνικών (Κεφάλαιο 3). Ο λόγος που επιβάλλει αυτή τη διαφοροποίηση είναι το γεγονός πως οι ευφυείς τεχνικές βασίζονται στην δημιουργία ενός μοντέλου κατά τη διάρκεια της εκπαίδευσης, το οποίο και αποτελεί τη βάση στην οποία γίνονται οι προβλέψεις.

Οι γραφοθεωρητικές τεχνικές πρόβλεψης εξετάζουν την δομή του γράφου και λαμβάνοντας υπόψη βασικά χαρακτηριστικά των κόμβων, όπως τον αριθμό των γειτόνων, τον αριθμό των κοινών γειτόνων μεταξύ δυο κόμβων ή τα μονοπάτια μεταξύ τους, υπολογίζουν έναν δείκτη ομοιότητας απαραίτητο για την διαδικασία πρόβλεψης νέων ακμών. Η διαδικασία υπολογισμού του δείκτη ομοιότητας μεταξύ δυο κόμβων μπορεί να είναι απλούστερη ή/και πιο σύνθετη.

Η χρήση ευφυών τεχνικών για την επίλυση του προβλήματος πρόβλεψης νέων ακμών στα μέσα κοινωνικής δικτύωσης απαιτεί την εκπαίδευση ταξινομητών. Βασική διαφορά σε σχέση με τις γραφοθεωρητικές τεχνικές είναι ότι η εκπαίδευση των ταξινομητών χωρίζεται στην φάση μάθησης και την φάση επικύρωσης. Η διαδικασία μάθησης των ταξινομητών είναι περίπλοκη και απαιτεί την μελέτη όλων των δεδομένων εισόδου και όχι μέρος αυτών όπως στις γραφοθεωρητικές τεχνικές (λ.χ. η τεχνική των κοινών γειτόνων που παρουσιάστηκε στο Κεφάλαιο 2 απαιτεί για τον προσδιορισμό της πληροφορία μόνο για τους κοινούς γείτονες δύο κόμβων x και y και όχι για όλο το δίκτυο). Πιο συγκεκριμένα, η δημιουργία του μοντέλου προβλέψεων κατά την χρήση ευφυών τεχνικών απαιτεί τον προσδιορισμό σημαντικών παραμέτρων του μοντέλου που υπολογίζονται από την συνολική μελέτη των δεδομένων εισόδου του δικτύου. Η φάση της επικύρωσης είναι πολύ σημαντική και απαραίτητη για την ποιοτική αξιολόγηση του μοντέλου.

Οι διαφορές της διαδικασίας υπολογισμού του βαθμού ομοιότητας μεταξύ δυο κόμβων του δικτύου, όπως αυτές περιγράφηκαν παραπάνω, είναι και η βασική αιτία ύπαρξης διαφορετικών πειραματικών πρωτοκόλλων μεταξύ των απλών και ευφυών τεχνικών.

4.1 Σύνολα δεδομένων

Στην παρούσα διπλωματική εργασία, χρησιμοποιήθηκαν δυο *σύνολα δεδομένων* (datasets) εισόδου, με την βοήθεια των οποίων δοκιμάστηκαν οι διάφοροι αλγόριθμοι πρόβλεψης ακμών. Πρόκειται για τα `soc-hamsterster` [Ross13b] και `soc-advogato` [Ross13a]. Και τα δύο σύνολα δεδομένων έχουν συλλεχθεί από τους Ryan A. Rossi και Nesreen K. Ahmed [Ross15].

Το πρώτο προέρχεται από την πλατφόρμα κοινωνικής δικτύωσης Hamsterster, η οποία έχει πλέον αναστείλει τη λειτουργία της. Αφορά τις σχέσεις φιλίας μεταξύ των χρηστών της εν λόγω πλατφόρμας, με τις ακμές του δικτύου να είναι μη κατευθυντικές και χωρίς βάρη. Οι κόμβοι του γράφου υποδηλώνουν τους χρήστες της ιστοσελίδας και οι ακμές τις σχέσεις φιλίας μεταξύ τους.

Χαρακτηριστικό	`soc-hamsterster`	`soc-advogato`
Κόμβοι	2,4K	5,2K
Ακμές	16,6K	47,3K
Πυκνότητα	0,57 %	0,35 %
Μέγιστος βαθμός	273	947
Ελάχιστον βαθμός	1	1
Μέσος βαθμός	13	18
Αριθμός τριγώνων	159,8K	499,7K
Μέσος αριθμός τριγώνων	65	96
Μέγιστος βαθμός τριγώνων	2,7K	11,4K
Μέσος συντελεστής συσταδοποίησης	0,537533	0,286832

Πίνακας 4.1: Χαρακτηριστικά συνόλων δεδομένων

Το δεύτερο σύνολο δεδομένων προέρχεται από την πλατφόρμα κοινωνικής δικτύωσης Advogato [Levi99]. Ο γράφος του συγκεκριμένου κοινωνικού δικτύου είναι μη κατευθυντικός και με βάρη στις ακμές. Στα πλαίσια αυτής της εργασίας, χρησιμοποιούμε μια απλουστευμένη εκδοχή του δικτύου, χωρίς την χρήση των βαρών στις ακμές. Όπως και στο προηγούμενο σύνολο δεδομένων, οι κόμβοι του γράφου αναφέρονται στους χρήστες του δικτύου και οι ακμές στις σχέσεις φιλίας μεταξύ των χρηστών. Ο Πίνακας 4.1 συνοψίζει τα κύρια χαρακτηριστικά των δύο συνόλων δεδομένων που χρησιμοποιήθηκαν στο πειραματικό κομμάτι της παρούσας εργασίας.

Παρατηρούμε ότι το σύνολο δεδομένων `soc-hamsterster` έχει μικρότερες διαστάσεις από το σύνολο δεδομένων `soc-advogato`. Συγκεκριμένα, το δεύτερο σύνολο έχει διπλάσιο αριθμό κόμβων και τριπλάσιο αριθμό ακμών σε σχέση με το πρώτο σύνολο. Επίσης, το πρώτο σύνολο είναι πιο πυκνό από το δεύτερο όπως μας δείχνει ο δείκτης πυκνότητας του Πίνακα 4.1.

Είναι επίσης ενδιαφέρον να σημειωθεί, πως και τα δυο σύνολα εμφανίζουν χαρακτηριστικά δικτύων ελεύθερης κλίμακας (Ενότητα 1.4). Βασική ιδιαιτερότητα των συγκεκριμένων δικτύων είναι ότι η κατανομή του βαθμού των κόμβων ακολουθεί ασυμπτωτικά την εκθετική συνάρτηση. Αυτό σημαίνει ότι ένας μικρός αριθμός κόμβων θα έχει υπερβολικά μεγαλύτερο βαθμό από τον μέσο βαθμό κόμβων του δικτύου. Αυτοί οι κόμβοι αποτελούν τα *κέντρα* (hubs) του δικτύου, ενώ η πλειοψηφία των κόμβων έχουν σχετικά χαμηλό βαθμό.

Χαρακτηριστική γραφική παράσταση των δικτύων ελεύθερης κλίμακας αποτελεί η γραμμική μείωση του λογαρίθμου του βαθμού του κόμβου ως προς τον λογάριθμο του πλήθους των κόμβων που έχουν αυτό τον βαθμό. Αυτή η μείωση είναι υπαρκτή και για τα δύο σύνολα δεδομένων που μελετώνται (Σχήμα 1.5a και 1.5b).

4.2 Πειραματικό πρωτόκολλο

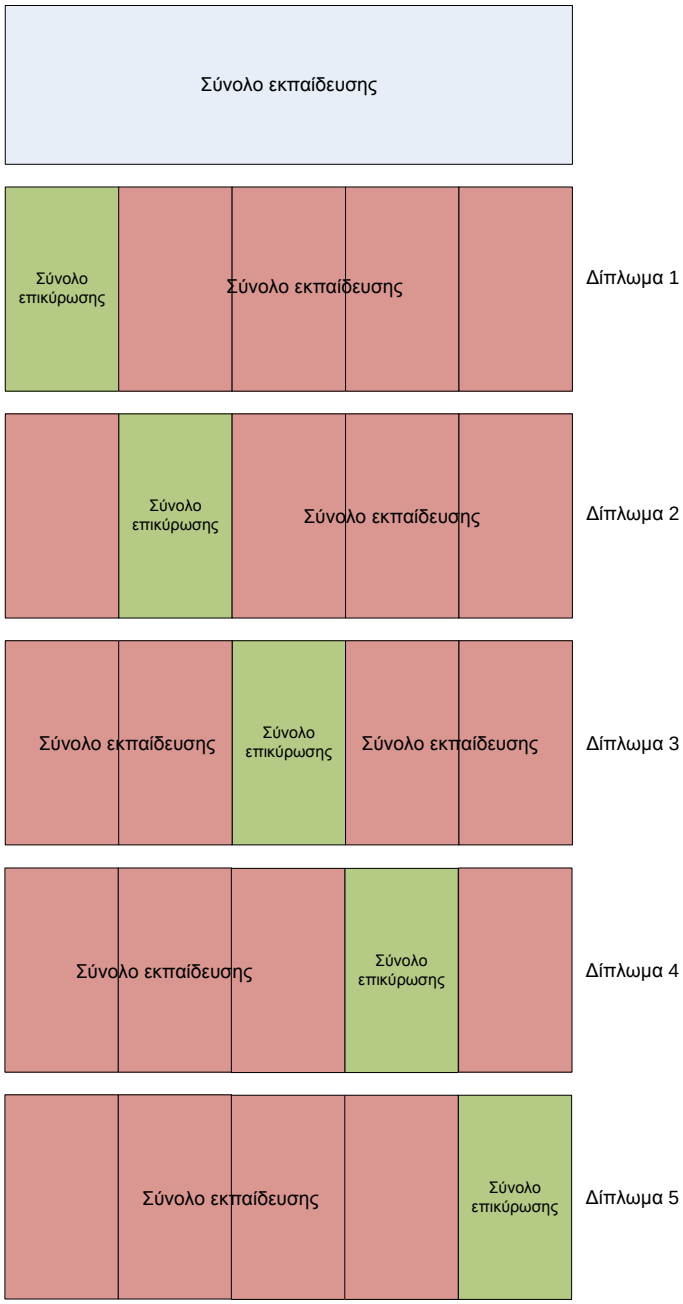
4.2.1 Μέθοδος διασταυρούμενης αντεπικύρωσης

Η μέθοδος της *διασταυρούμενης αντεπικύρωσης* (cross validation) χρησιμοποιείται κυρίως σε προβλήματα προβλέψεων και είναι μια μέθοδος επικύρωσης μοντέλων. Η επίλυση τέτοιων προβλημάτων έχει δυο στάδια. Το στάδιο της εκπαίδευσης του μοντέλου προβλέψεων και το στάδιο ελέγχου, όπου γίνεται η κατηγοριοποίηση νέων δειγμάτων. Συνήθως, τα δεδομένα του προβλήματος αποτελούν ένα σύνολο δειγμάτων, το οποίο χρησιμοποιείται για την εκπαίδευση. Είναι χρήσιμο, να μπορεί να υπολογιστεί η απόδοση του μοντέλου από την φάση της εκπαίδευσης, πριν ακόμα η διαδικασία καταλήξει στο στάδιο του ελέγχου. Αυτό επιτυγχάνεται με την μέθοδο της διασταυρούμενης αντεπικύρωσης.

Το σύνολο δειγμάτων εκπαίδευσης χωρίζεται κατάλληλα σε ένα ή περισσότερα υποσύνολα. Σε κάθε φάση της εκπαίδευσης ένα υποσύνολο θα αποτελεί το *σύνολο επαλήθευσης* (validation set) και όλα τα υπόλοιπα υποσύνολα θα δημιουργούν το *σύνολο εκπαίδευσης* (training set). Το σύνολο επαλήθευσης λειτουργεί όπως και το *σύνολο ελέγχου* (test set) αλλά στο στάδιο της εκπαίδευσης. Με

αυτόν τον τρόπο, προβλήματα όπως η *υπερπροσαρμογή* (overfitting) μπορούν να αντιμετωπιστούν πολύ νωρίς κατά την διαδικασία της μάθησης.

Μια από τις πιο γνωστές μεθόδους αντεπικύρωσης είναι η μέθοδος *διασταυρούμενης αντεπικύρωσης k-διπλωμάτων* (k-fold cross validation). Σε αυτήν, το σύνολο εκπαίδευσης χωρίζεται σε k ίσα υποσύνολα τυχαία. Σε κάθε βήμα της μεθόδου τα $k-1$ υποσύνολα χρησιμοποιούνται για την εκπαίδευση του μοντέλου και τα δεδομένα του τελευταίου υποσυνόλου χρησιμοποιούνται για την επικύρωση του μοντέλου. Η διαδικασία αυτή επαναλαμβάνεται k φορές. Κάθε φορά ένα διαφορετικό υποσύνολο χρησιμοποιείται ως σύνολο επικύρωσης και όλα τα υπόλοιπα για την εκπαίδευση. Στο τέλος της διαδικασίας μπορεί να υπολογιστεί ένας μέσος όρος των μετρικών απόδοσης του μοντέλου. Το Σχήμα 4.1 βοηθά στην καλύτερη κατανόηση της διαδικασίας.



Σχήμα 4.1: Παράδειγμα μεθόδου διασταυρούμενης αντεπικύρωσης 5 διπλωμάτων.

Παρατηρείται ότι κάθε δείγμα του συνόλου εισόδου βρίσκεται σε σύνολο επικύρωσης ακριβώς μια φορά. Με την διαδικασία των k-διπλωμάτων και τον υπολογισμό του μέσου όρου των μετρικών

απόδοσης από κάθε επανάληψη, λαμβάνεται μια γενική εικόνα για το πως θα συμπεριφέρεται το μοντέλο στο στάδιο ελέγχου, μετά το τέλος της εκπαίδευσης του.

4.3 Πειραματικό πρωτόκολλο γραφοθεωρητικών τεχνικών

Αφού αναφερθήκαμε στη μέθοδο της διασταυρούμενης αντεπικύρωσης, σε αυτήν την ενότητα θα αναλυθεί το πειραματικό πρωτόκολλο των γραφοθεωρητικών τεχνικών.

Θεωρούμε έναν γράφο $G(V, E)$ όπου V είναι το σύνολο των κόμβων του γράφου και E είναι το σύνολο ακμών του γράφου, όπως αυτά δίνονται σαν δεδομένα εισόδου. Η επίλυση του προβλήματος πρόβλεψης ακμών θα γίνει σε αυτό το στιγμιότυπο του γράφου. Η βασική ιδέα είναι να αφαιρεθεί ένα μέρος των ακμών του υπάρχοντος γράφου και κατόπιν να γίνει προσπάθεια ανάκτησής του, μελετώντας τον υπόλοιπο γράφο του δικτύου.

Αρχικά, σε ένα σύνολο E_{exp} τοποθετούνται όλες οι ακμές που προσπίπτουν σε κόμβους με βαθμό μεγαλύτερο από ένα ελάχιστο κατώφλι d_t. Ο λόγος που τίθεται το εν λόγω κατώφλι είναι πως η αφαίρεση ακμών από κόμβους με χαμηλό βαθμό μπορεί να προκαλέσει την απομόνωσή τους από το υπόλοιπο δίκτυο και έτσι να επηρεαστεί η απόδοση των μετρικών. Το σύνολο ακμών E_{exp} θα αποτελέσει και το σύνολο εκπαίδευσης κατά την διαδικασία επίλυσης του προβλήματος. Οι ακμές που παραλείπονται, δηλαδή οι ακμές που ο ένας τουλάχιστον κόμβος στα άκρα τους έχει βαθμό μικρότερο του d_t, σχηματίζουν το σύνολο E_{excl}. Από διάφορα πειράματα που πραγματοποιήθηκαν και στις δύο συλλογές δεδομένων που εξετάζονται, βρέθηκε ότι η καταλληλότερη τιμή για το d_t είναι 4.

Στην συνέχεια, εκτελείται η μέθοδος της διασταυρούμενης αντεπικύρωσης k-διπλωμάτων για το σύνολο E_{exp}. Τα δεδομένα του συνόλου χωρίζονται σε k ίσα υποσύνολα. Σε κάθε επανάληψη/δίπλωμα i, το υποσύνολο E^i_{val} είναι το σύνολο ελέγχου και όλα τα υπόλοιπα $k-1$ υποσύνολα το σύνολο εκπαίδευσης E^i_{tr}. Όπως ειπώθηκε και πριν, κάθε ακμή του συνόλου E_{exp} θα βρεθεί σε σύνολο επικύρωσης ακριβώς μια φορά.

Σε κάθε επανάληψη της μεθόδου, αφαιρούνται όλες οι ακμές του συνόλου ελέγχου E^i_{val} από τον γράφο G. Το αποτέλεσμα αυτής της ενέργειας είναι η δημιουργία ενός νέου υπογράφου $G'(V, E^i_{\text{tr}})$, όπου V το σύνολο των κόμβων του αρχικού γράφου G και E^i_{tr} το σύνολο των ακμών εκπαίδευσης, το οποίο προκύπτει από την διαδικασία που περιγράφηκε ακριβώς παραπάνω. Τέλος, σε ένα σύνολο V^i_{val} αποθηκεύονται όλοι οι κόμβοι που είναι άκρα των ακμών του συνόλου ελέγχου E^i_{val}. Προφανώς ισχύει πως $V^i_{\text{val}} \in V$.

4.3.1 Διαδικασία πρόβλεψης ακμών με χρήση γραφοθεωρητικών τεχνικών και μέτρηση της απόδοσης

Κατά την διαδικασία της πρόβλεψης ακμών με χρήση των γραφοθεωρητικών τεχνικών, υπολογίζεται μια τιμή για κάθε ακμή που δημιουργείται από όλα τα πιθανά ζευγάρια κόμβων που βρίσκονται στο σύνολο V^i_{val}. Έστω ότι το σύνολο V^i_{val} περιέχει n κόμβους. Τότε θα πρέπει για κάθε κόμβο να ελεγχθούν $n-1$ ακμές, δηλαδή όλες οι πιθανές ακμές με όλους τους υπόλοιπους κόμβους του V^i_{val}. Στο τέλος αυτής της διαδικασίας, θα πρέπει να υπάρχει μια λίστα με $n-1$ ακμές για κάθε κόμβο του συνόλου V^i_{val}. Ανάμεσα σε όλες αυτές τις $n*(n-1)$ πιθανές ακμές βρίσκονται και οι ακμές του συνόλου ελέγχου E^i_{val} που υπάρχουν πραγματικά στον γράφο G.

Για τον υπολογισμό της πιθανότητας εμφάνισης της κάθε ακμής, εφαρμόζονται οι γραφοθεωρητικές τεχνικές που αναπτύχθηκαν στο Κεφάλαιο 2 και πιο συγκεκριμένα οι:

1. Τυχαίες προβλέψεις
2. Κοινοί γείτονες
3. Συντελεστής Jaccard
4. Συντελεστής Adamic / Adar

5. Προτιμώμενη προσκόλληση
6. Δείκτης Katz
7. Μετρική HittingTime
8. Μετρική CommuteTime
9. Κανονικοποιημένη μετρική HittingTime
10. Κανονικοποιημένη μετρική CommuteTime
11. Μετρική Rooted PageRank

Αφού κατασκευαστεί η λίστα που περιέχει όλες τις πιθανές ακμές προς όλους τους υπόλοιπους κόμβους του συνόλου V^i_{val} για κάθε υπό εξέταση κόμβο, αυτή ταξινομείται κατά φθίνουσα σειρά ανάλογα με την πιθανότητα εμφάνισης της κάθε ακμής, έτσι όπως αυτή αποτυπώνεται από την εκάστοτε τεχνική και συγκρατούνται οι m ακμές με την καλύτερη τιμή. Με αυτόν τον τρόπο, μπορούν να προβλεφθούν για κάθε υπό εξέταση κόμβο οι m πιο πιθανές ακμές.

Τέλος, ελέγχοντας τη λίστα με τις m καλύτερες ακμές κάθε κόμβου και το πόσες από αυτές βρίσκονται στο σύνολο ελέγχου E^i_{val}, υπολογίζεται η απόδοση της εκάστοτε τεχνικής για κάθε κόμβο ξεχωριστά, καθώς και ο μέσος όρος αυτής.

4.3.2 Ψευδοκώδικας πειραματικού πρωτοκόλλου γραφοθεωρητικών τεχνικών

Ο Αλγόριθμος 1 συνοψίζει το πειραματικό πρωτόκολλο που χρησιμοποιήθηκε στις γραφοθεωρητικές τεχνικές. Όπως φαίνεται και από τον ψευδοκώδικα που παρατίθεται, το συγκεκριμένο πρωτόκολλο προϋποθέτει την ύπαρξη συνάρτησης *Split(E_{exp}, k)* στη γραμμή 8, η οποία λαμβάνει ως ορίσματα το σύνολο των ακμών εκπαίδευσης E_{exp} και το πλήθος των διπλωμάτων k. Αφού χωρίσει το σύνολο E_{exp} σε k ισομεγέθη υποσύνολα, τα επιστρέφει στον πίνακα E_{val}.

Στη γραμμή 19, η συνάρτηση *ComputeScore($G(V, E^i_{tr}), (u, v)$)* λαμβάνει ως όρισμα τον γράφο καθώς ένα ζεύγος κόμβων και υλοποιεί μια από τις γραφοθεωρητικές τεχνικές που αναφέρθηκαν στην Ενότητα 4.3.1. Η επιστρεφόμενη τιμή $score$ είναι επί της ουσίας η μέτρηση του δείκτη ομοιότητας μεταξύ των δύο κόμβων για την εκάστοτε γραφοθεωρητική τεχνική που εξετάζεται.

Στη γραμμή 24, η συνάρτηση *KeepBestEdges($scoredEdges, u, m$)* λαμβάνει ως όρισμα τη δομή $scoredEdges$ (η οποία περιέχει ζεύγη κόμβων και τις αντίστοιχες τιμές ομοιότητας που έχουν υπολογιστεί για αυτά) , έναν κόμβο/χρήστη u καθώς και ένα πλήθος ακμών m. Επιστρέφει τις m πιθανότερες να εμφανιστούν ακμές, οι οποίες έχουν αφετηρία τον χρήστη/κόμβο u.

Τέλος στη γραμμή 26, η συνάρτηση *PerformanceMetrics($bestEdges, E^i_{val}$)* λαμβάνει ως όρισμα τη δομή $bestEdges$, που περιέχει τις πιθανότερες να εμφανιστούν ακμές, καθώς και τις ακμές του συνόλου E^i_{val}, δηλαδή τις ακμές που πραγματικά υπάρχουν. Συγκρίνοντας αυτά τα δύο σύνολα, υπολογίζεται η απόδοση της εκάστοτε γραφοθεωρητικής τεχνικής.

4.4 Πειραματικό πρωτόκολλο ευφυών τεχνικών

Όπως περιγράφηκε και στο Κεφάλαιο 3, η επίλυση του προβλήματος της πρόβλεψης ακμών από τις ευφυείς τεχνικές πραγματοποιείται με τη χρήση ταξινομητών. Όπως φάνηκε από την ανάλυση, υπάρχουν πολλά είδη ταξινομητών που λειτουργούν και επιλύουν το πρόβλημα με διαφορετικούς τρόπους. Το κοινό στοιχείο όλων αυτών των ταξινομητών είναι η είσοδος που λαμβάνουν και η έξοδος που παράγουν.

Βασική λειτουργία των ταξινομητών είναι η δημιουργία ενός μοντέλου προβλέψεων εξετάζοντας τα δείγματα του συνόλου εκπαίδευσης. Όταν τελειώσει η φάση μάθησης και το μοντέλο έχει πλέον δημιουργηθεί, ο ταξινομητής προσπαθεί να προβλέψει αν ένα ζεύγος κόμβων θα συνδεθεί μελλοντικά

Αλγόριθμος 1 Πειραματικό πρωτόκολλο γραφοθεωρητικών τεχνικών

Require: γράφος $G(V, E)$, πλήθος διπλωμάτων k, πλήθος βέλτιστων ακμών m

function GraphTheoreticalExperiment(G, k, m)
  $E_{\text{exp}} \leftarrow \emptyset$ ▷ Το σύνολο ακμών της πειραματικής διαδικασίας
  for each $(u, v) \in E$ **do** ▷ Εξαίρεση από την πειραματική διαδικασία των ακμών...
    if $d_u \geq 4$ **and** $d_v \geq 4$ **then** ▷ ...που προσπίπτουν σε κόμβους με βαθμό μικρότερο του 4
      $E_{\text{exp}} \leftarrow E_{\text{exp}} \cup (u, v)$
    end if
  end for
  $E_{\text{val}} \leftarrow \text{Split}(E_{\text{exp}}, k)$ ▷ Ο πίνακας E_{val} περιέχει τα k διαφορετικά υποσύνολα του E_{exp}
  $i \leftarrow 1$
  while $i \leq k$ **do**
    $V^i_{\text{val}} \leftarrow \emptyset$ ▷ Το i-οστό σύνολο κόμβων επαλήθευσης
    $E^i_{\text{val}} \leftarrow E_{\text{val}}[i]$ ▷ Το i-οστό σύνολο ακμών επαλήθευσης
    $E^i_{\text{tr}} \leftarrow E \setminus E^i_{\text{val}}$ ▷ Το i-οστό σύνολο ακμών εκπαίδευσης
    for each $(u, v) \in E^i_{\text{val}}$ **do**
      $V^i_{\text{val}} \leftarrow V^i_{\text{val}} \cup u \cup v$
    end for
    $scoredEdges \leftarrow \emptyset$
    for each $u, v \in V^i_{\text{val}}, u \neq v$ **do** ▷ Για κάθε ζεύγος κόμβων στο i-οστό σύνολο επαλήθευσης
      $score \leftarrow \text{ComputeScore}(G(V, E^i_{\text{tr}}), (u, v))$
      $scoredEdges \leftarrow scoredEdges \cup ((u, v), score)$
    end for
    $bestEdges \leftarrow \emptyset$
    for each $u \in V^i_{\text{val}}$ **do**
      $bestEdges[u] \leftarrow \text{KeepBestEdges}(scoredEdges, u, m)$
    end for
    $Performance[i] \leftarrow \text{PerformanceMetrics}(bestEdges, E^i_{\text{val}})$
    $i \leftarrow i + 1$
  end while
end function

με ακμή ή όχι. Με άλλα λόγια, η έξοδος του ταξινομητή είναι μια ετικέτα με θετική τιμή «1» ή αρνητική τιμή «0» για κάθε εξεταζόμενο ζεύγος κόμβων.

Το σύνολο εκπαίδευσης του ταξινομητή αποτελείται από ένα σύνολο διανυσμάτων χαρακτηριστικών, ένα διάνυσμα για κάθε πιθανό ζεύγος κόμβων του δικτύου, είτε αυτό ανήκει στην κλάση «0», δηλαδή δεν συνδέεται με ακμή, είτε ανήκει στην κλάση «1», δηλαδή συνδέεται με ακμή. Έστω δίκτυο $G(V, E)$ όπου V το σύνολο των κόμβων και E το σύνολο των ακμών. Έστω επίσης, το σύνολο E_{all}, που περιέχει όλα τα πιθανά ζεύγη κόμβων του V. Για κάθε ένα ζεύγος στο σύνολο E_{all} σχηματίζεται ένα διάνυσμα χαρακτηριστικών, το οποίο αποτελείται από τα παρακάτω στοιχεία:

1. Τους κόμβους που βρίσκονται στα άκρα της ακμής
2. Την ετικέτα της κατηγορίας στην οποία ανήκει η ακμή. Η ετικέτα έχει τιμή «1» για τις ακμές που υπάρχουν στο δίκτυο και τιμή «0» για αυτές που δεν υπάρχουν.
3. Τον βαθμό και των δυο κόμβων που βρίσκονται στα άκρα της ακμής
4. Την τιμή του συντελεστή Adamic-Adar για αυτήν την ακμή
5. Την τιμή του συντελεστή των Κοινών Γειτόνων για αυτήν την ακμή

6. Την τιμή του δείκτη Katz χωρίς βάρη, υπολογισμένο για όλα τα μονοπάτια μεταξύ των δυο κόμβων

7. Την τιμή του κανονικοποιημένου δείκτη CommuteTime

8. Την τιμή του PageRank

Όπως είναι αναμενόμενο, τα ζεύγη κόμβων του συνόλου E_{all} που βρίσκονται στην κατηγορία «0» θα είναι πολύ περισσότερα από τα ζεύγη κόμβων της κατηγορίας «1». Ειδικότερα, στα μέσα κοινωνικής δικτύωσης που βασικό χαρακτηριστικό τους είναι η αραιότητα, αναμένεται οι υπαρκτές ακμές να αποτελούν λιγότερο από το 1% όλων των πιθανών ακμών μεταξύ των κόμβων του δικτύου. Η συγκεκριμένη παρατήρηση μπορεί να επιβεβαιωθεί και διαισθητικά, μιας και αν επιλεγούν εντελώς τυχαία δύο χρήστες, το πιο πιθανό είναι να μην γνωρίζονται μεταξύ τους. Κάτω από αυτές τις συνθήκες ανισορροπίας των δύο κλάσεων, η εκπαίδευση ενός ταξινομητή είναι πάρα πολύ δύσκολη.

Για την επίλυση αυτού του προβλήματος, αναζητήθηκε κάποιος δείκτης/συντελεστής ο οποίος θα περιόριζε το πλήθος των υποψήφιων ακμών προς εξέταση, χωρίς ωστόσο να αλλοιώσει σε μεγάλο βαθμό τα χαρακτηριστικά του συνόλου δεδομένων και κυριότερα την αναλογία μεταξύ των μελών των δύο κλάσεων. Μετά από ενδελεχή μελέτη και πειραματική επαλήθευση, καταλληλότερος αποδείχτηκε ο συντελεστής Adamic-Adar. Αφού ορίστηκε ένα ελάχιστο κατώφλι για τον εν λόγω συντελεστή, όσα ζεύγη κόμβων είχαν τιμή μικρότερη από αυτό αφαιρέθηκαν από το σύνολο E_{all}, σχηματίζοντας κατ' αυτό τον τρόπο το νέο σύνολο E_{th}.

Ο βασικός στόχος αυτού του βήματος είναι να συγκρατηθούν όσες περισσότερες υπαρκτές ακμές του δικτύου (ακμές της κατηγορίας «1») μέσα στο σύνολο E_{th}, με το ποσοστό αυτών των ακμών να φτάσει 5% έως 10% των συνολικών ακμών. Με αυτόν τον τρόπο, η εκπαίδευση ενός ταξινομητή είναι πλέον εφικτή, ενώ παράλληλα δεν αλλοιώνονται τα χαρακτηριστικά του προβλήματος σε σημαντικό βαθμό, καθώς οι υπαρκτές ακμές του δικτύου αποτελούν και πάλι ένα μικρό μέρος του συνόλου εκπαίδευσης. Σε αυτό το σημείο πρέπει να τονιστεί ότι στην Ενότητα 3.5.3 παρουσιάστηκε μια λύση για το πρόβλημα της ανισορροπίας κλάσεων σε ένα πρόβλημα δυαδικής κατηγοριοποίησης, με την επιβολή βαρών στις κλάσεις. Το Σχήμα 4.2 βοηθά στην καλύτερη κατανόηση της παραπάνω διαδικασίας.

4.4.1 Διαδικασία εκπαίδευσης του μοντέλου προβλέψεων και μέτρηση της απόδοσης του

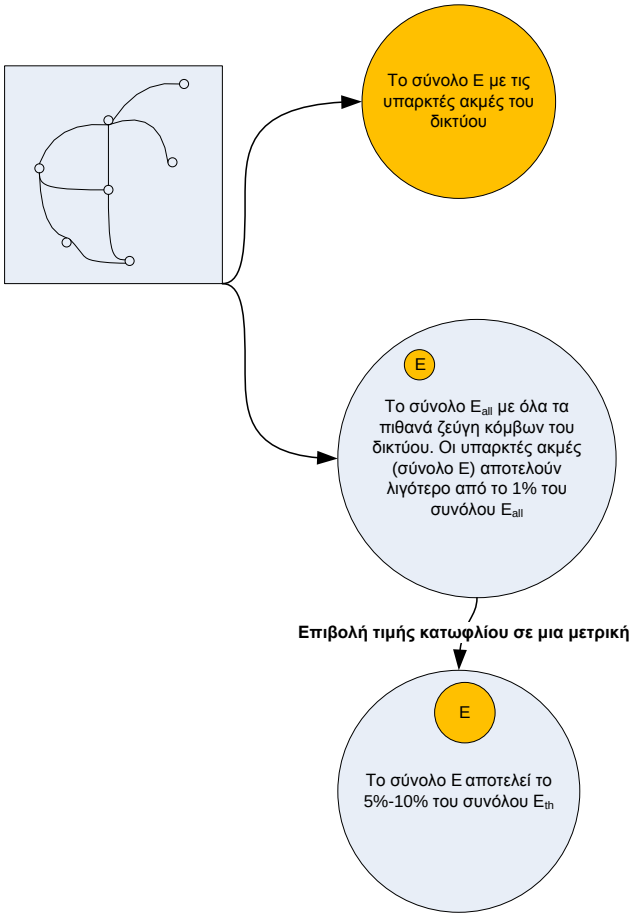
Στην προηγούμενη Ενότητα περιγράφηκε η διαδικασία δημιουργίας των δεδομένων εισόδου του ταξινομητή. Η εκπαίδευση του μοντέλου και η μέτρηση της απόδοσης του γίνεται και σε αυτή την περίπτωση με τη μέθοδο της διασταυρούμενης αντεπικύρωσης k-διπλωμάτων.

Το σύνολο E_{th} χωρίζεται σε k υποσύνολα. Σε κάθε επανάληψη της μεθόδου, ένα υποσύνολο χρησιμοποιείται σαν σύνολο επικύρωσης του μοντέλου και τα υπόλοιπα $k-1$ υποσύνολα ως το σύνολο εκπαίδευσης. Η εκπαίδευση του μοντέλου γίνεται με επίβλεψη, που σημαίνει ότι ο ταξινομητής εξετάζει το διάνυσμα χαρακτηριστικών κάθε ζεύγους κόμβων στο σύνολο εκπαίδευσης καθώς και την ετικέτα της κλάσης τους.

Αφού τελειώσει η φάση της μάθησης και το μοντέλο προβλέψεων έχει δημιουργηθεί, στην συνέχεια εισάγονται ως είσοδος στο μοντέλο τα διανύσματα χαρακτηριστικών κάθε ζεύγους κόμβων που βρίσκονται στο σύνολο επικύρωσης και λαμβάνεται ως έξοδος η ετικέτα της κατηγορίας που ανήκει η ακμή, όπως αυτή προβλέφτηκε από το μοντέλο.

Στην επόμενη επανάληψη της μεθόδου διασταυρούμενης αντεπικύρωσης, το σύνολο επικύρωσης αλλάζει και η εκπαίδευση του μοντέλου ξεκινά από την αρχή.

Γνωρίζοντας την πραγματική κατηγορία που ανήκουν τα ζεύγη κόμβων του συνόλου επικύρωσης και έχοντας λάβει την πρόβλεψη του μοντέλου για αυτά, μπορούν να υπολογιστούν διάφορες μετρικές απόδοσης του μοντέλου. Στα πλαίσια αυτής της εργασίας, υπολογίζεται η *Ακρίβεια* (Precision), η *Ανάκληση* (Recall) και η μετρική F1 για όλο το σύνολο επικύρωσης, αλλά και για την κατηγορία «1» ξεχωριστά (Ενότητα 4.4.3). Ουσιαστικά, για το πρόβλημα της πρόβλεψης ακμών στα μέσα κοινωνικής



Σχήμα 4.2: Προεπεξεργασία δεδομένων εισόδου των ταξινομητών κατά την μελέτη των ευφυών τεχνικών για την επίλυση του προβλήματος πρόβλεψης ακμών στα κοινωνικά δίκτυα

δικτύωσης ενδιαφερόμαστε κυρίως για τα ζεύγη κόμβων της κατηγορίας «1», καθώς μεταξύ αυτών είναι πιθανό να αναπτυχθούν μελλοντικά ακμές.

4.4.2 Ψευδοκώδικας πειραματικού πρωτοκόλλου ευφυών τεχνικών

Ο Αλγόριθμος 2 συνοψίζει το πειραματικό πρωτόκολλο που χρησιμοποιήθηκε στις ευφυείς τεχνικές. Αρχικά η συνάρτηση *ComputeAttributeVector(*$G, (u, v)$*)* (γραμμή 9), υπολογίζει το διάνυσμα χαρακτηριστικών για κάθε ζεύγος κόμβων του γράφου. Επίσης, και αυτή την περίπτωση, η συνάρτηση *Split(*E_{exp}, k*)* στη γραμμή 9 λαμβάνει ως ορίσματα το σύνολο των ακμών εκπαίδευσης πάνω από ένα κατώφλι t για τον συντελεστή Adamic-Adar καθώς και το πλήθος των διπλωμάτων k και τα χωρίζει σε k ισομεγέθη υποσύνολα, τα οποία επιστρέφονται στη δομή E_{val}.

Στην γραμμή 18, η συνάρτηση *TrainClassifier(*G, E^i_{tr}*)* κατασκευάζει και εκπαιδεύει το εκάστοτε μοντέλο ταξινομητή , το οποίο επιστρέφεται στη δομή $model$. Στην αμέσως επόμενη γραμμή, καλείται η συνάρτηση *Predict(*E^i_{test}*)* του μοντέλου, με όρισμα τις ακμές στο σύνολο ελέγχου E^i_{test}. Αφού πραγματοποιηθούν οι προβλέψεις, επιστρέφεται η δομή $predictions$, που περιέχει την κάθε ακμή του συνόλου ελέγχου μαζί με την πρόβλεψη για την κλάση (0 ή 1) στην οποία ανήκει.

Τέλος, η συνάρτηση *ComputePerformance(*$predictions, E^i_{val}$*)* (γραμμή 20) υπολογίζει την απόδοση του ταξινομητή, συγκρίνοντας τις προβλέψεις του εκπαιδευμένου μοντέλου και την πραγματική κλάση που ανήκουν οι ακμές του συνόλου επικύρωσης.

Αλγόριθμος 2 Πειραματικό πρωτόκολλο ευφυών τεχνικών

Require: γράφος $G(V, E)$, πλήθος διπλωμάτων k, τιμή κατωφλίου t

procedure MachineLearningExperiment(G, k, t)
  $E_{\text{th}} \leftarrow \emptyset$
  for each $u, v \in V, u \neq v$ **do**
    $AttributeVector \leftarrow$ ComputeAttributeVector($G, (u, v)$)
    if $AttributeVector.AdamicAdar \geq t$ **then**
      $E_{\text{th}}[(u, v)] \leftarrow AttributeVector$
    end if
  end for
  $E_{\text{val}} \leftarrow$ Split(E_{th}, k) ▷ Η δομή E_{val} περιέχει τα k διαφορετικά υποσύνολα του E_{th}
  $i \leftarrow 1$
  while $i \leq k$ **do**
    $E^i_{\text{val}} \leftarrow E_{\text{val}}[i]$ ▷ Το i-οστό σύνολο ακμών επικύρωσης
    $E^i_{\text{test}} \leftarrow \emptyset$ ▷ Το i-οστό σύνολο κόμβων ελέγχου
    for each $(u, v) \in E^i_{\text{val}}, u \neq v$ **do**
      $E^i_{\text{test}} \leftarrow E^i_{\text{test}} \cup (u, v)$
    end for
    $E^i_{\text{tr}} \leftarrow E_{\text{val}} \setminus E^i_{\text{val}}$ ▷ Το i-οστό σύνολο κόμβων εκπαίδευσης
    $model \leftarrow$ TrainClassifier(G, E^i_{tr})
    $predictions \leftarrow model$.predict($E^i_{\text{test}}$)
    $Performance_i \leftarrow$ ComputePerformance($predictions, E^i_{\text{val}}$)
    $i \leftarrow i + 1$
  end while
end procedure

4.4.3 Ορισμός μετρικών απόδοσης ευφυών τεχνικών

Σε προβλήματα δυαδικής κατηγοριοποίησης οι μετρικές που χρησιμοποιούνται για την μέτρηση της απόδοσης των ταξινομητών είναι αυτές της ακρίβειας, της ανάκλησης και του δείκτη F1.

Η ακρίβεια ορίζεται ως το ποσοστό των σωστών προβλέψεων σε σχέση με τον αριθμό όλων των προβλέψεων που έκανε ο ταξινομητής. Η ανάκληση ορίζεται ως το ποσοστό των σωστών προβλέψεων σε σχέση με τον συνολικό αριθμό δειγμάτων που έπρεπε να έχουν προβλεφθεί σωστά.

Στο συγκεκριμένο πρόβλημα δυαδικής κατηγοριοποίησης που μελετούμε, το πρόβλημα της πρόβλεψης ακμών, μας ενδιαφέρει ο υπολογισμός της απόδοσης ως προς την κλάση «1» του προβλήματος και μόνο. Δεν μας ενδιαφέρει να προβλέψουμε ζεύγη κόμβων που ανήκουν στην κλάση «0», καθώς αυτοί δεν πρόκειται να συνδεθούν μεταξύ τους με ακμή στο μέλλον. Έτσι, οι ορισμοί της ακρίβειας και της ανάκλησης μπορούν να προσαρμοστούν στο πρόβλημα μας.

Η ακρίβεια μπορεί να οριστεί ως το ποσοστό των ζευγών κόμβων που ο ταξινομητής προέβλεψε ότι σωστά ανήκουν στην κλάση «1» σε σχέση με τον συνολικό αριθμό των ζευγών κόμβων που ο ταξινομητής προέβλεψε ότι ανήκουν στην κλάση «1», είτε αυτοί ανήκαν πραγματικά στην κλάση «1» είτε ανήκαν στην κλάση «0». Σύμφωνα με το Σχήμα 4.3, η ακρίβεια προκύπτει από τον λόγο του κελιού (1,1) ως προς το άθροισμα των κελιών της πρώτης στήλης,

$$\text{Ακρίβεια} = \frac{\text{Αληθώς Θετικό}}{\text{Αληθώς Θετικό + Ψευδώς Θετικό}} \tag{4.1}$$

Παρομοίως, η ανάκληση μπορεί να οριστεί ως το ποσοστό των ζευγών κόμβων που ο ταξινομητής προέβλεψε σωστά ότι ανήκουν στην κλάση «1» σε σχέση με τον συνολικό αριθμό των ζευγών του συνόλου επικύρωσης που ανήκουν στην κλάση «1». Με άλλα λόγια, η ανάκληση δηλώνει το ποσοστό των ακμών της κλάσης «1» που βρήκε ο ταξινομητής εξετάζοντας το σύνολο επικύρωσης.

Σχήμα 4.3: Ορισμός ακρίβειας και ανάκλησης

Η ανάκληση προκύπτει από τον λόγο του κελιού (1,1) ως προς το άθροισμα των κελιών της πρώτης γραμμής (Σχήμα 4.3),

$$\text{Ανάκληση} = \frac{\text{Αληθώς Θετικό}}{\text{Αληθώς Θετικό + Ψευδώς Αρνητικό}} \tag{4.2}$$

Τέλος, η μετρική $F1$ συνδυάζει την ακρίβεια και την ανάκληση, υπολογίζοντας τον αρμονικό μέσο όρο τους.

$$F1 = 2 * \frac{\text{Ακρίβεια * Ανάκληση}}{\text{Ακρίβεια + Ανάκληση}} \tag{4.3}$$

Κεφάλαιο 5

Αποτελέσματα

Σε αυτό το Κεφάλαιο παρουσιάζονται τα ευρήματα της έρευνάς μας πάνω στο πρόβλημα της πρόβλεψης ακμών στα μέσα κοινωνικής δικτύωσης. Χρησιμοποιήθηκαν οι συλλογές δεδομένων `soc-hamsterster` και `soc-advogato` που παρουσιάστηκαν με περισσότερες λεπτομέρειες στην Ενότητα 4.1 και υλοποιήθηκαν τόσο *γραφοθεωρητικές τεχνικές* (Κεφάλαιο 2) όσο και *ευφυείς τεχνικές* (Κεφάλαιο 3).

Πιο συγκεκριμένα, από τις *τεχνικές βασισμένες στους γειτόνους* (Ενότητα 2.2.1) παρουσιάζονται αποτελέσματα για τις μεθόδους των *κοινών γειτόνων* (Εξίσωση 2.1), του *συντελεστή Jaccard* (Εξίσωση 2.2) καθώς και του *συντελεστή Adamic-Adar* (Εξίσωση 2.3). Ο λόγος που έγινε η συγκεκριμένη επιλογή είναι πως οι συγκεκριμένες μέθοδοι παρουσίαζαν τα καλύτερα αποτελέσματα και στις δύο συλλογές δεδομένων που εξετάστηκαν, σε σύγκριση με τις ομοειδείς τεχνικές που παρουσιάστηκαν στην σχετική ενότητα (Ενότητα 2.2.1). Ακολουθώντας αντίστοιχη συλλογιστική, από τις *τεχνικές βασισμένες στη διαδρομή* (Ενότητα 2.2.2) παρουσιάζονται αποτελέσματα για τον *δείκτη Katz* (Εξίσωση 2.13) και τέλος από τις *τεχνικές βασισμένες στους τυχαίους περιπάτους* (Ενότητα 2.2.3) παρουσιάζονται αποτελέσματα για τους δείκτες *hitting time* (Εξίσωση 2.20), *commute time* (Εξίσωση 2.22) καθώς και για τις κανονικοποιημένες εκδοχές τους (Εξισώσεις 2.27 και 2.28 αντίστοιχα)

Από τις ευφυείς τεχνικές παρουσιάζονται αποτελέσματα για τον *απλό μπεϋζιανό ταξινομητή* (Ενότητα 3.2), τον ταξινομητή των *k-πλησιέστερων γειτόνων* (Ενότητα 3.3), τα *δέντρα αποφάσεων* (Ενότητα 3.4), τις *μηχανές διανυσμάτων υποστήριξης* (Ενότητα 3.5) καθώς και για τη μέθοδο της *στοχαστικής καθόδου κλίσης* (Ενότητα 3.6). Τέλος, για την συνολικότερη εκτίμηση της συνεισφοράς της κάθε τεχνικές (γραφοθεωρητική ή ευφυούς) επιλέχθηκε ως *σύστημα αναφοράς* (reference system ή baseline) ένα σύστημα τυχαίων προβλέψεων, το οποίο δηλαδή συνέδεε με ακμή δύο κόμβους με ομοιόμορφα τυχαίο τρόπο.

Στην συνέχεια παρουσιάζονται οι γραφικές παραστάσεις της απόδοσης των τεχνικών πρόβλεψης για κάθε σύνολο δεδομένων και σε μορφή τέτοια ώστε η σύγκριση των διαφόρων τεχνικών να γίνεται όσο το δυνατόν πιο εύκολη. Επίσης, για τις γραφοθεωρητικές τεχνικές περιλαμβάνονται γραφικές παραστάσεις για διαφορετικό αριθμό διπλωμάτων της μεθόδου διασταυρούμενης αντεπικύρωσης.

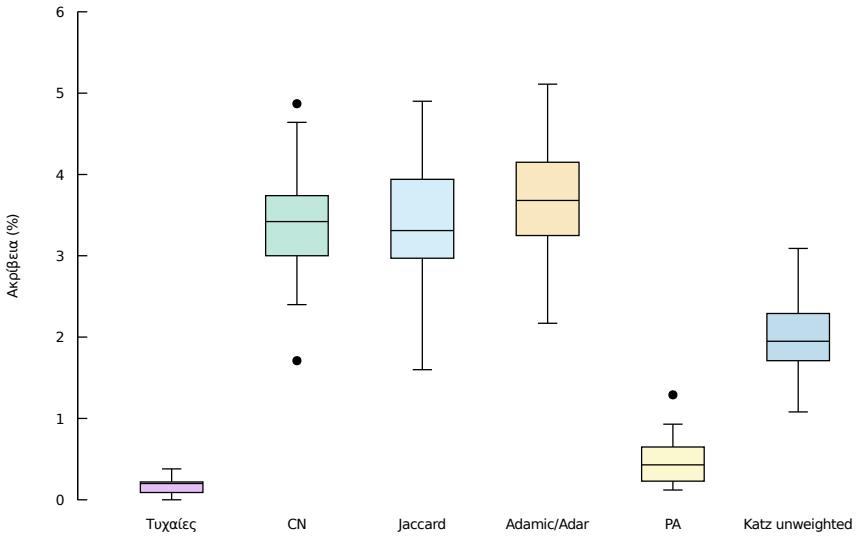
5.1 Απόδοση γραφοθεωρητικών τεχνικών

5.1.1 Τεχνικές βασισμένες στους γειτόνους και τη διαδρομή

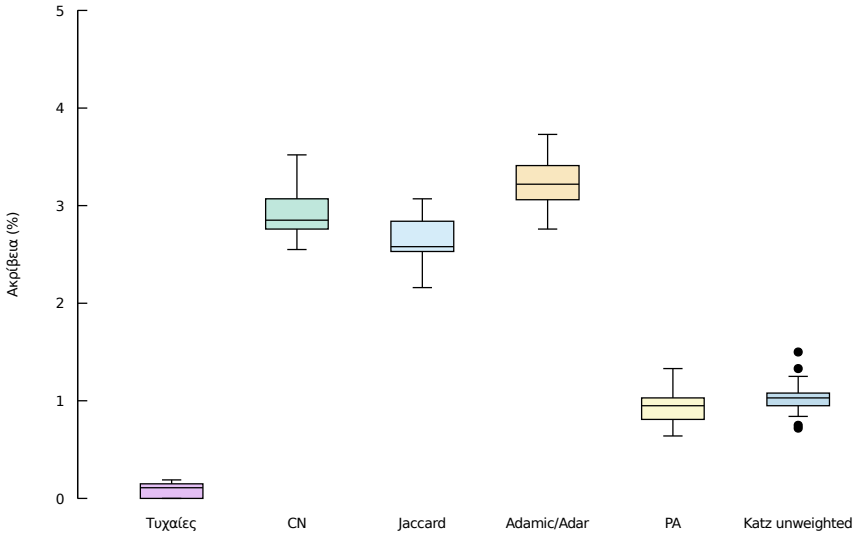
Στα Σχήματα 5.1 και 5.2 παρουσιάζονται τα αποτελέσματα των τεχνικών βασισμένων στους γειτόνους και τη διαδρομή για τις μεθόδους διασταυρούμενης αντεπικύρωσης 25 και 100 διπλωμάτων αντίστοιχα, υπό μορφή *θηκογράμματος* (boxplot). Η πρώτη παρατήρηση που έχουμε να κάνουμε είναι πως όλες οι τεχνικές εμφανίζουν την ίδια μέση τιμή στην ίδια συλλογή δεδομένων και για τις δύο μεθόδους διασταυρούμενης αντεπικύρωσης. Η μόνη διαφορά έγκειται στην διασπορά των μετρήσεων, η οποία είναι μεγαλύτερη για τη διασταυρούμενη αντεπικύρωση 100 διπλωμάτων σε σύγκριση με την αντίστοιχη των 25 διπλωμάτων.

Η αύξηση αυτή της διασποράς ερμηνεύεται ως εξής: από τη στιγμή που η συλλογή των δεδομένων έχει σταθερό μέγεθος, η αύξηση του αριθμού των διπλωμάτων έχει ως αποτέλεσμα τη δημιουργία μικρότερων συνόλων ελέγχου (Κεφάλαιο 4). Συνεπώς, ένα σφάλμα ταξινόμησης ενός ζεύγους

κόμβων στο μικρότερο σύνολο ελέγχου αναμένεται να έχει μεγαλύτερη βαρύτητα στη διαμόρφωση της μέσης τιμής της αντίστοιχης μετρικής για το συγκεκριμένο δίπλωμα απ' ότι ένα σφάλμα σε μεγαλύτερο σύνολο ελέγχου. Άρα, οι αποκλίσεις στη μέση τιμή μεταξύ των μετρικών απόδοσης των συνόλων ελέγχου θα είναι μεγαλύτερες, πράγμα που αποτυπώνεται στην αύξηση της διασποράς των θηκογραμμάτων.



(a) Συλλογή δεδομένων `soc-hamsterster`



(b) Συλλογή δεδομένων `soc-advogato`

Σχήμα 5.1: Απόδοση γραφοθεωρητικών τεχνικών βασισμένων στους γειτόνους και τη διαδρομή (διασταυρωμένη αντεπικύρωση 25 διπλωμάτων)

Ας υποτεθεί για παράδειγμα, ένα σύνολο με 100 δείγματα. Αν αρχικά εφαρμοστεί η μέθοδος διασταυρούμενης αντεπικύρωσης 10 διπλωμάτων, σε κάθε δίπλωμα τα 90 δείγματα θα αποτελούν το σύνολο εκπαίδευσης και τα 10 υπολειπόμενα δείγματα το σύνολο ελέγχου. Η λανθασμένη ταξινόμηση ενός δείγματος από το σύνολο ελέγχου, θα επιφέρει μείωση κατά 10% στην απόδοση του εκάστοτε ταξινομητή. Σε αντίθετη περίπτωση, αν γίνει εφαρμογή της μεθόδου διασταυρούμενης αντεπικύρωσης 50 διπλωμάτων, όπου σε κάθε δίπλωμα το σύνολο ελέγχου αποτελείται από μόλις 2 δείγματα, η λανθασμένη ταξινόμηση ενός από αυτά θα επιφέρει μείωση κατά 50% στην απόδοση του ταξινομητή. Γίνεται σαφές, πως το σφάλμα ταξινόμησης αποκτά μεγαλύτερη βαρύτητα καθώς το μέγεθος

του συνόλου ελέγχου μειώνεται.

Η ίδια ερμηνεία μπορεί να δοθεί ως εξήγηση και για την διαφορά που υπάρχει στην διασπορά των δυο συνόλων δεδομένων. Παρατηρείται ότι το σύνολο `soc-hamsterter` έχει μεγαλύτερη διασπορά σε σχέση με το σύνολο `soc-advogato`. Από τον Πίνακα 4.1 φαίνεται ότι το πρώτο σύνολο έχει λιγότερες ακμές και επομένως το σύνολο ελέγχου για κάθε επανάληψη της μεθόδου διασταυρούμενης αντεπικύρωσης k-διπλωμάτων θα περιέχει λιγότερα ζεύγη κόμβων προς ταξινόμηση σε σχέση με το αντίστοιχο σύνολο ελέγχου του `soc-advogato` και για ίδιο αριθμό διπλωμάτων.

Παρά την ύπαρξη μεγαλύτερης διασποράς, το σύνολο δεδομένων `soc-hamsterster` φαίνεται να έχει οριακά καλύτερη απόδοση σχεδόν σε όλες τις γραφοθεωρητικές τεχνικές που εξετάστηκαν. Όπως αναφέρθηκε στο Κεφάλαιο 2 οι γραφοθεωρητικές τεχνικές στηρίζονται κυρίως στην δομή του δικτύου και η απόδοσή τους επηρεάζεται σημαντικά από τα χαρακτηριστικά του. Ένα από τα χαρακτηριστικά που φαίνεται να έχει ιδιαίτερη σημασία είναι η πυκνότητα του δικτύου. Όπως παρατηρείται στον Πίνακα 4.1 το σύνολο `soc-hamsterster` έχει συντελεστή πυκνότητας ελαφρώς μεγαλύτερο από τον αντίστοιχο του συνόλου `soc-advogato` και αυτός είναι ο λόγος της οριακής βελτίωσης στην απόδοση των τεχνικών.

Σε γενικές γραμμές παρατηρείται ότι όλες οι τεχνικές που είναι βασισμένες στους γειτόνους έχουν αρκετά χαμηλή απόδοση και στις δύο συλλογές δεδομένων (κάτω του 10%), δηλαδή γίνεται σωστή πρόβλεψη για ελάχιστες από τις ακμές που βρίσκονται στο σύνολο ελέγχου.

Όπως αναμενόταν, οι τυχαίες προβλέψεις έχουν την χαμηλότερη ακρίβεια από όλες τις τεχνικές, μιας και τα μέσα κοινωνικής δικτύωσης ανήκουν στην κατηγορία των αραιών δικτύων (Ενότητα 1.4). Ένα δίκτυο/γράφος $G(V, E)$ θεωρείται αραιό αν το πλήθος των ακμών του είναι περίπου της ίδιας τάξης μεγέθους με το πλήθος των κόμβων του (ισχύει δηλαδή $|V| \sim |E|$) και σε κάθε περίπτωση είναι πολύ μικρότερο από το πλήθος των πιθανών συνδυασμών των κόμβων του ανά ζεύγη

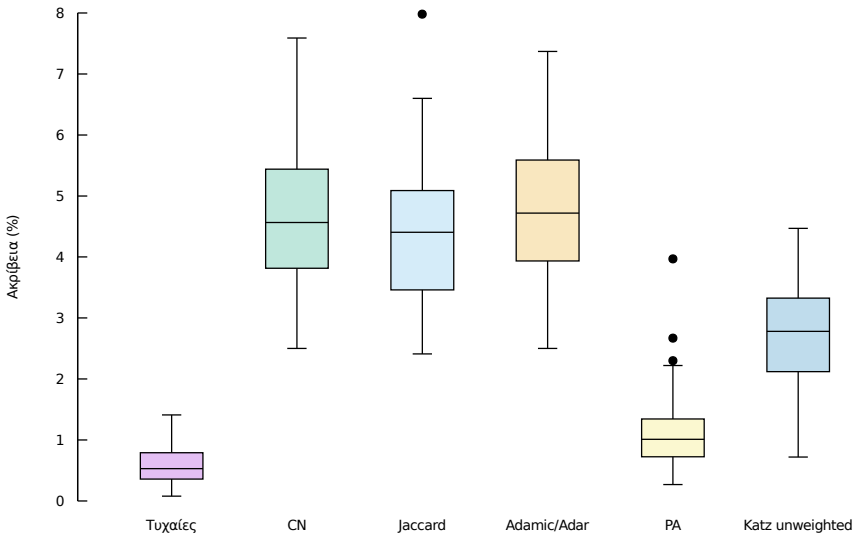
$$\binom{V}{2} = \frac{|V|(|V|-1)}{2} \approx \frac{1}{2}|V|^2 \gg |E| \tag{5.1}$$

Η Εξίσωση 5.1 ισχύει και για τις δύο συλλογές δεδομένων που εξετάστηκαν, μιας και από τον Πίνακα 4.1 προκύπτει για το `soc-hamsterster` $2,88 \times 10^6 \gg 16,6 \times 10^3$ και για το `soc-advogato` $13,52 \times 10^6 \gg 47,3 \times 10^3$. Συνεπώς, η πιθανότητα προσθήκης νέας ακμής ομοιόμορφα τυχαία μεταξύ των κόμβων του G υπολογίζεται ως εξής

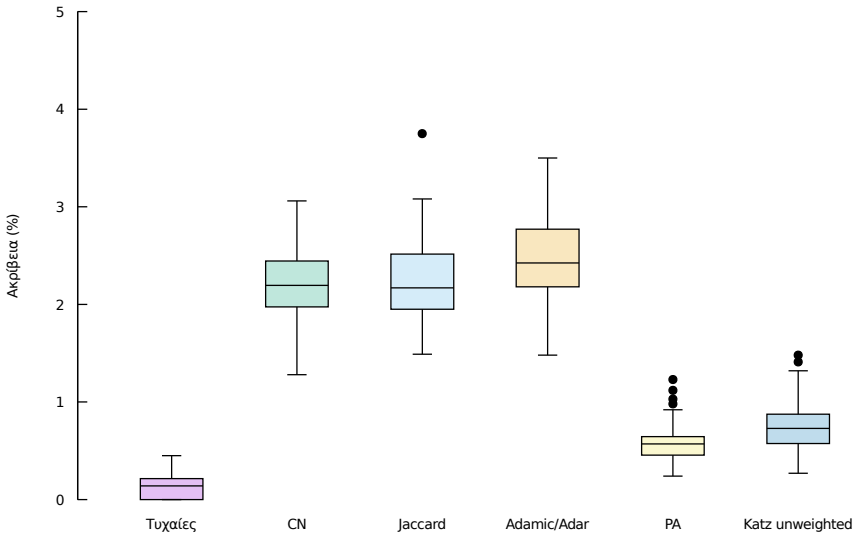
$$p_{\text{new}} = \frac{1}{\binom{V}{2} - |E|} = \frac{2}{|V|(|V|-1) - 2|E|} \approx \frac{2}{|V|^2} \tag{5.2}$$

Στις περιπτώσεις των συλλογών δεδομένων που εξετάζονται, η πιθανότητα της Εξίσωσης 5.2 γίνεται $0,347 \times 10^{-6}$ για το `soc-hamsterster` και $0,074 \times 10^{-6}$ για το `soc-advogato`. Παρατηρούμε δηλαδή ότι η συγκεκριμένη πιθανότητα είναι υπερβολικά μικρή, πράγμα που μπορεί να επιβεβαιωθεί και διαισθητικά, μιας και η πιθανότητα να γνωρίζονται μεταξύ τους δύο εντελώς τυχαίοι χρήστες ενός μέσου κοινωνικής δικτύωσης είναι επίσης πάρα πολύ μικρή. Επίσης, η εν λόγω πιθανότητα είναι αντιστρόφως ανάλογη του τετραγώνου του πλήθους των κόμβων του δικτύου, πράγμα το όποιο δικαιολογεί το γεγονός πως είναι μεγαλύτερη στο `soc-hamsterster` απ' ότι στο `soc-advogato`. Η συγκεκριμένη παρατήρηση επαληθεύεται και πειραματικά, μιας και το σύστημα τυχαίων προβλέψεων εμφανίζει οριακά καλύτερη απόδοση στην πρώτη συλλογή δεδομένων απ' ότι στη δεύτερη.

Μια ακόμα σημαντική παράμετρος στην μελέτη των τυχαίων προβλέψεων είναι η καλή γνώση των χαρακτηριστικών των συνόλων δεδομένων που εξετάζονται. Για παράδειγμα, ας θεωρηθούν δυο κοινωνικά δίκτυα όπου αναπαριστούν μια ακαδημαϊκή κοινότητα, στα οποία οι κόμβοι του γράφου αποτελούν τους καθηγητές / ερευνητές της κοινότητας και οι ακμές αναπαριστούν τυχόν συνεργασίες μεταξύ των κόμβων για την δημοσίευση μιας ερευνητικής εργασίας. Στο ένα κοινωνικό δίκτυο οι ερευνητές ανήκουν όλοι σε ένα και μόνο ερευνητικό πεδίο, ενώ στο δεύτερο οι ερευνητές ανήκουν σε διαφορετικά ερευνητικά πεδία. Μια τυχαία πρόβλεψη ζεύγους κόμβων στο πρώτο δίκτυο έχει μεγαλύτερες πιθανότητες να αποτελέσει μελλοντική ακμή του δικτύου, καθώς οι συνεργασίες μεταξύ ερευνητών του ίδιου επιστημονικού πεδίου είναι πιο πιθανές να συμβούν.



(a) Συλλογή δεδομένων `soc-hamsterster`



(b) Συλλογή δεδομένων `soc-advogato`

Σχήμα 5.2: Απόδοση γραφοθεωρητικών τεχνικών βασισμένων στους γειτόνους και τη διαδρομή (διασταυρωμένη αντεπικύρωση 100 διπλωμάτων)

Όπως φαίνεται στα Σχήματα 5.1 και 5.2 οι τρεις τεχνικές που φαίνεται να έχουν την καλύτερη απόδοση και στα δυο σύνολα δεδομένων είναι οι κοινοί γείτονες, ο συντελεστής Jaccard και ο συντελεστής Adamic-Adar. Η ομοιότητα στην απόδοση αυτών των τεχνικών, μπορεί να εξηγηθεί από τον τρόπο με τον οποίο υπολογίζουν τον βαθμό ομοιότητας μεταξύ δυο κόμβων (Εξισώσεις 2.1, 2.2 και 2.3 αντίστοιχα). Το κοινό σημείο των τριών προαναφερόμενων τεχνικών είναι ο υπολογισμός του αριθμού των κοινών γειτόνων μεταξύ δυο κόμβων.

Μια επιπλέον παρατήρηση που πρέπει να γίνει, είναι ότι ο συντελεστής Jaccard παρουσιάζει μεγαλύτερη διασπορά σε σχέση με τις άλλες δυο τεχνικές. Παρά την ομοιότητα των τριών τεχνικών, ο συντελεστής Jaccard είναι η μόνη τεχνική που κανονικοποιεί τον βαθμό ομοιότητας ως προς το άθροισμα των γειτόνων των δυο κόμβων (Εξίσωση 2.2). Αυτή η προσέγγιση είναι προβληματική για κόμβους κοινωνικών δικτύων, λόγω των χαρακτηριστικών δικτύων ελεύθερης κλίμακας που αυτά εμφανίζουν (Ενότητα 1.4). Συγκεκριμένα, η ύπαρξη ελαχίστων δημοφιλών κόμβων με πολύ μεγάλο βαθμό, προκαλεί τη δραματική μείωση του συντελεστή και πιθανώς κακή απόδοση. Είναι επίσης

ενδιαφέρον, να μελετηθεί διαισθητικά η περίπτωση υπολογισμού του βαθμού ομοιότητας δυο δημοφιλών κόμβων ενός δικτύου (λ.χ. δυο πολιτικοί) με χρήση του συντελεστή Jaccard. Το πιθανότερο ενδεχόμενο είναι ότι οι δυο δημοφιλείς κόμβοι θα ενώνονται με κάποιο τρόπο μεταξύ τους, παρότι ο συντελεστής θα επιστρέψει μια πολύ μικρή τιμή ομοιότητας λόγω του υψηλού βαθμού των κόμβων.

Τα χαρακτηριστικά των δικτύων ελεύθερης κλίμακας είναι και η αιτία που ο συντελεστής προτιμώμενης προσκόλλησης έχει πολύ χαμηλή απόδοση. Ο συντελεστής δίνει μεγάλη βαρύτητα στις ακμές που προσπίπτουν σε δημοφιλείς κόμβου του δικτύου (Εξίσωση 2.4). Ο ελάχιστος αριθμός δημοφιλών κόμβων στα κοινωνικά δίκτυα σε συνδυασμό με το γεγονός ότι η επιλογή των ζευγών κόμβων που αποτελούν το σύνολο ελέγχου γίνεται εντελώς τυχαία (Ενότητα 4.3.1), καθορίζουν την χαμηλή απόδοση του συντελεστή.

Ο δείκτης Katz χωρίς βάρη υπολογίζει τον βαθμό ομοιότητας δυο κόμβων με βάση τον αριθμό των μονοπατιών που ενώνουν τους δυο κόμβους (Εξίσωση 2.13). Λόγω της αραιότητας των κοινωνικών δικτύων τα διαθέσιμα προς εξερεύνηση μονοπάτια του γράφου είναι περιορισμένα και συνεπώς, η απόδοση τεχνικών που βασίζονται στην εύρεση μονοπατιών είναι αρκετά χαμηλή. Χαρακτηριστικά που δηλώνουν την αραιότητα των δικτύων των συνόλων δεδομένων που χρησιμοποιήσαμε μπορούν να φανούν στον Πίνακα 4.1.

Συμπερασματικά, παρατηρείται ότι η μελέτη της δομής του δικτύου μπορεί να βοηθήσει στην πρόβλεψη μελλοντικών ακμών έναντι της τυχαίας επιλογής. Αυτό μπορεί να φανεί στα Σχήματα 5.1 και 5.2 όπου για τα δυο σύνολα δεδομένων που μελετήθηκαν οι γραφοθεωρητικές τεχνικές υπερτερούν. Παρ' όλα αυτά η βελτίωση της απόδοσης δεν είναι σημαντική, καθώς οι γραφοθεωρητικές τεχνικές φαίνεται ότι επηρεάζονται πολύ από τα χαρακτηριστικά δικτύων ελεύθερης κλίμακας που εμφανίζουν τα κοινωνικά δίκτυα. Συγκεκριμένα, η αραιότητα και η μεγάλη ανισορροπία στους βαθμούς των κόμβων αποτελούν δυο πολύ σημαντικές αιτίες για την χαμηλή απόδοση αυτών των τεχνικών. Αξίζει να σημειωθεί όμως, ότι η συμπεριφορά των γραφοθεωρητικών τεχνικών είναι σχεδόν όμοια και για τα δυο σύνολα δεδομένων.

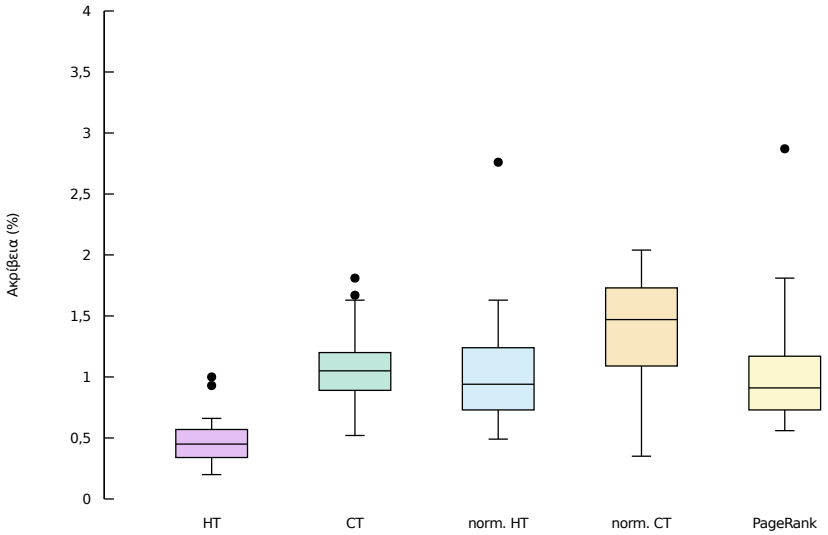
5.1.2 Τεχνικές βασισμένες στους τυχαίους περιπάτους

Στα Σχήματα 5.3 και 5.4 παρουσιάζονται τα αποτελέσματα των τεχνικών που βασίζονται στους τυχαίους περιπάτους για τις μεθόδους διασταυρούμενης αντεπικύρωσης 25 και 100 διπλωμάτων αντίστοιχα. Τα αποτελέσματα παρουσιάζονται σε μορφή θηκογράμματος.

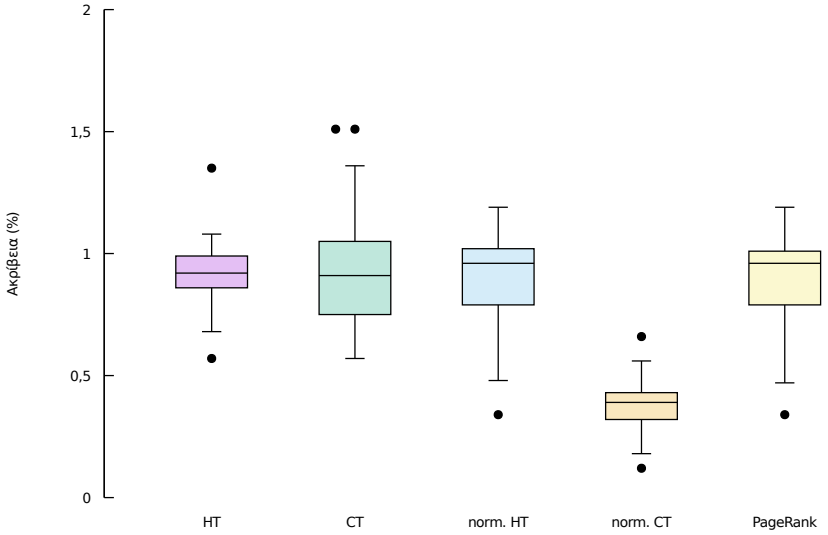
Και σε αυτήν την περίπτωση, η αραιότητα των δικτύων που μελετήθηκαν φαίνεται πως διαδραματίζει κυρίαρχο ρόλο στην πολύ χαμηλή απόδοση των τεχνικών αυτών. Η έλλειψη πολλών εναλλακτικών μονοπατιών μεταξύ των κόμβων του δικτύου και η τυχαιότητα στην οποία βασίζεται η έννοια του τυχαίου περιπάτου, συγκρατούν την απόδοση όλων των τεχνικών που εξετάζονται σε αυτήν την ενότητα σε επίπεδα κάτω του 5%.

Μια ακόμα σημαντική παράμετρος που επηρεάζει την απόδοση των τυχαίων περιπάτων είναι η ύπαρξη πολλών απομονωμένων νησίδων, των οποίων το σύνολο σχηματίζει τελικά το γενικότερο δίκτυο. Στα περισσότερα κοινωνικά δίκτυα οι άνθρωποι οργανώνονται σε κλειστές ομάδες, όπου οι σχέσεις μεταξύ των μελών της ομάδας είναι πάρα πολλές αλλά η επικοινωνία με το υπόλοιπο δίκτυο είναι ασθενής ή και πολλές φορές ανύπαρκτη. Από γραφοθεωρητική σκοπιά, η επικοινωνία μεταξύ των απομονωμένων δομικών στοιχείων του γράφου γίνεται με ελάχιστες (ή και καμιά) ακμές. Η μελέτη των τυχαίων περιπάτων σε ένα τέτοιο δίκτυο είναι πρακτικά πολύ δύσκολη, αφού η προσπέλαση των νησίδων είναι πολλές φορές αδύνατη.

Παρ' όλα αυτά, οι εξεταζόμενες τεχνικές εμφανίζουν μεταξύ τους μικρές αλλά εμφανείς διαφορές. Για παράδειγμα, στο `soc-hamsterster` ο δείκτης Commute Time εμφανίζει διπλάσια απόδοση από τον δείκτη Hitting Time (Σχήματα 5.3a και 5.4a). Αν ληφθούν υπόψη οι Εξισώσεις 2.21 και 2.22, φαίνεται ότι ο CT προκύπτει από τον HT, υπολογίζοντας τον αναμενόμενο αριθμό βημάτων από τον κόμβο x στον κόμβο y και πίσω ξανά στον x. Με άλλα λόγια, ο δείκτης CT εξετάζει την διπλάσια πληροφορία σε σύγκριση με τον HT, πράγμα το οποίο επηρεάζει την ικανότητα πρόβλεψης της εν λόγω τεχνικής. Αυτή η βελτίωση «χάνεται» ωστόσο, στο πιο αραιό δίκτυο `soc-advogato`, όπου είναι μεγαλύτερο το πλήθος των κόμβων που δεν συνδέονται μεταξύ τους με κανένα μονοπάτι, όποτε σε



(a) Συλλογή δεδομένων `soc-hamsterster`



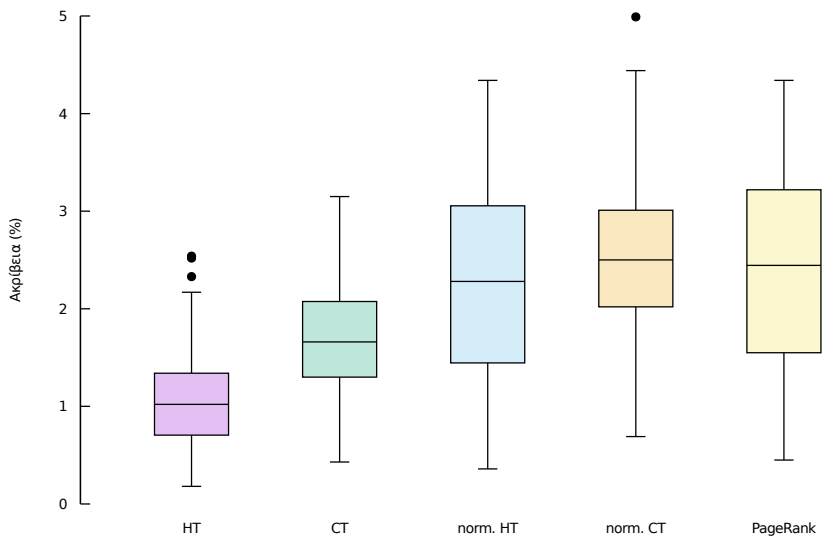
(b) Συλλογή δεδομένων `soc-advogato`

Σχήμα 5.3: Απόδοση τεχνικών βασισμένων στους τυχαίους περιπάτους (διασταυρούμενη αντεπικύρωση 25 διπλωμάτων)

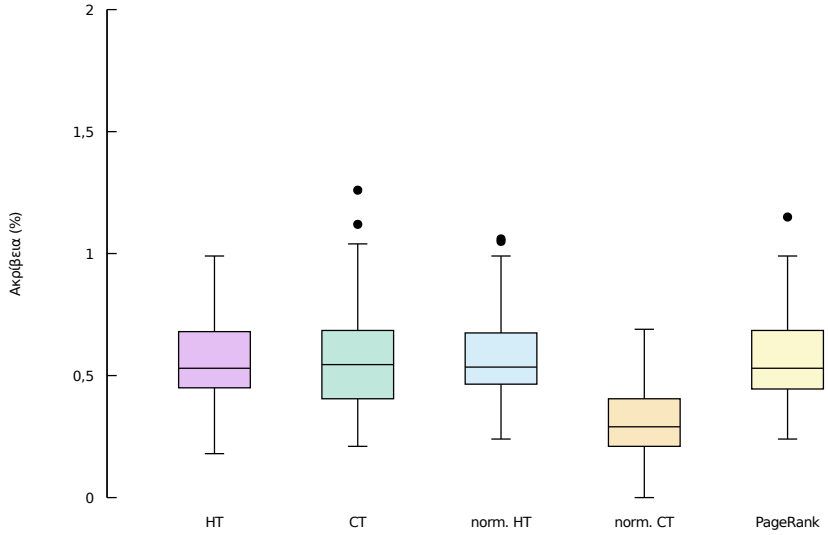
αυτή την περίπτωση ο CT δεν μπορεί να προσφέρει κάτι παραπάνω από τον HT, με αποτέλεσμα να εμφανίζουν σχεδόν την ίδια απόδοση (Σχήματα 5.3b και 5.4b).

Στην αραιότητα του δικτύου μπορεί να αποδοθεί και η χειροτέρευση της απόδοσης της τεχνικής του κανονικοποιημένου Commute Time στο `soc-advogato` σε σύγκριση με το `soc-hamsterster`. Από την Εξίσωση 2.28 φαίνεται ότι η εν λόγω κανονικοποίηση προκύπτει από τον πολλαπλασιασμό των τιμών του δείκτη Hitting Time με την πιθανότητα ένας τυχαίος περίπατος να βρεθεί στους κόμβους x και y στην στατική κατανομή (Εξίσωση 2.17). Όπως όμως αναφέρθηκε και πρωτύτερα, οι συγκεκριμένες πιθανότητες μικραίνουν όσο η αραιότητα του δικτύου μεγαλώνει, μιας και τα μονοπάτια μεταξύ των κόμβων είτε ελαχιστοποιούνται είτε διαρρηγνύονται τελείως.

Τέλος αξίζει να σημειωθεί ότι η απόδοση του PageRank είναι παρόμοια και στα δύο δίκτυα που εξετάστηκαν (γύρω στο 1%) παρότι το ένα έχει διπλάσια πυκνότητα από το άλλο. Και σε αυτή την περίπτωση το γεγονός ερμηνεύεται με την ύπαρξη των πολλών απομονωμένων νησίδων που δεν επικοινωνούν μεταξύ τους, μια πραγματικότητα που δεν επιτρέπει στον συγκεκριμένο αλγόριθμο να



(a) Συλλογή δεδομένων `soc-hamsterster`



(b) Συλλογή δεδομένων `soc-advogato`

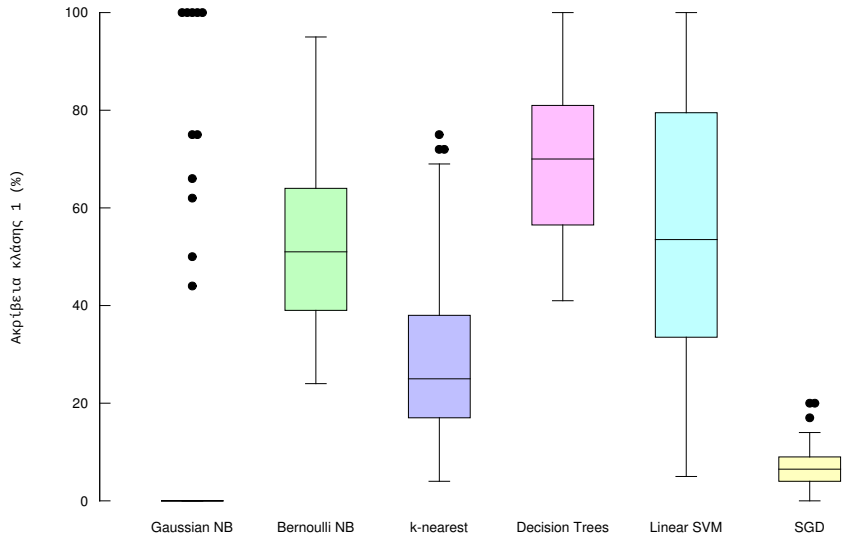
Σχήμα 5.4: Απόδοση τεχνικών βασισμένων στους τυχαίους περιπάτους (διασταυρούμενη αντεπικύρωση 100 διπλωμάτων)

ξεδιπλώσει την πλήρη δυναμική του.

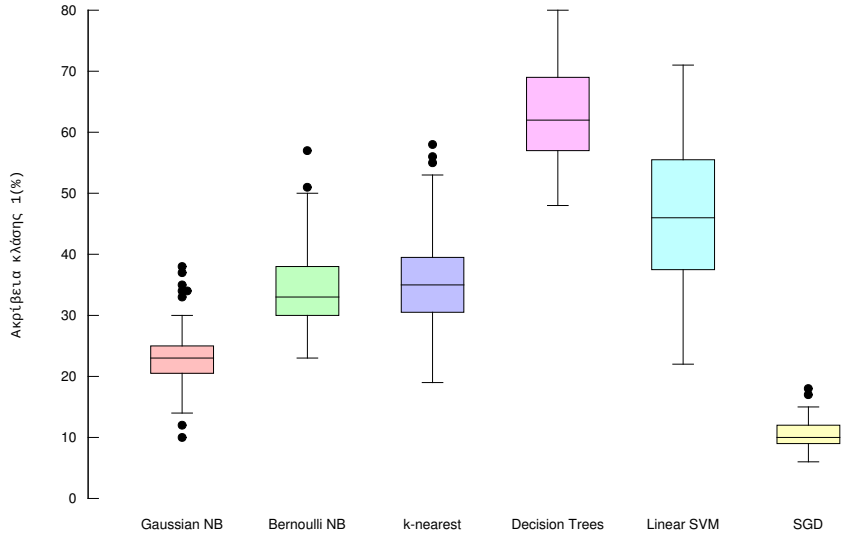
5.2 Απόδοση ευφυών τεχνικών

Σε αυτήν την ενότητα παρουσιάζεται, με την βοήθεια γραφικών παραστάσεων, η απόδοση των ευφυών τεχνικών επίλυσης του προβλήματος πρόβλεψης ακμών στα μέσα κοινωνικής δικτύωσης, για τους ταξινομητές που εξετάστηκαν λεπτομερώς στο Κεφάλαιο 3 της παρούσας εργασίας. Επίσης, γίνεται αναφορά στο σύνολο των υπερπαραμέτρων που μελετήθηκαν για κάθε ταξινομητή και σημειώνονται οι βέλτιστες τιμές που επιλέχθηκαν.

Οι γραφικές παραστάσεις της απόδοσης των ταξινομητών αναφέρονται σε 100 επαναλήψεις της μεθόδου διασταυρούμενης αντεπικύρωσης και απεικονίζουν τις τρεις μετρικές απόδοσης που παρουσιάστηκαν στην Ενότητα 4.4.3; την *Ακρίβεια*, την *Ανάκληση* και τον δείκτη *F1*. Τα μεγέθη αυτά δίνονται μόνο για την κλάση 1 (θετική κλάση) του προβλήματος, αφού από την κλάση αυτή θα προκύψουν



(a) Συλλογή δεδομένων `soc-hamsterster`



(b) Συλλογή δεδομένων `soc-advogato`

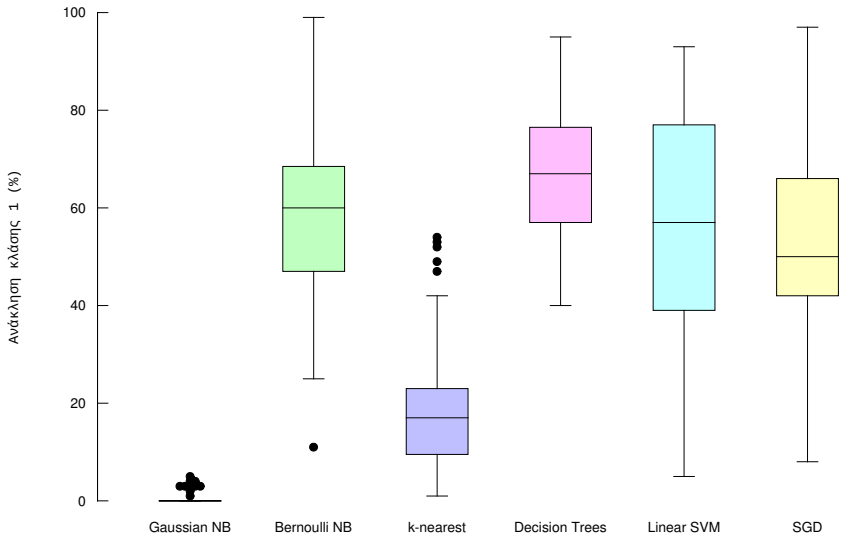
Σχήμα 5.5: Ακρίβεια ευφυών τεχνικών για την θετική κλάση (διασταυρούμενη αντεπικύρωση 100 διπλωμάτων)

οι τυχόν μελλοντικές ακμές του δικτύου.

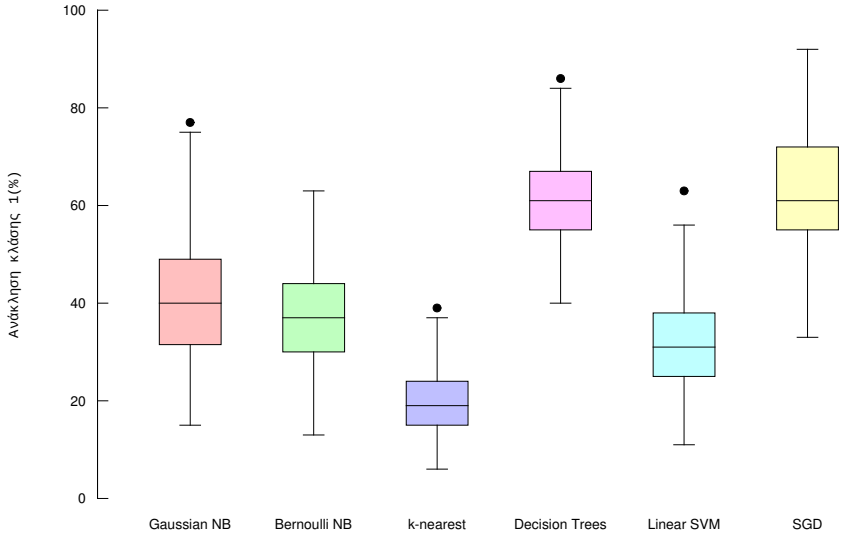
5.2.1 Παράμετροι

Για τον απλό μπεϋζιανό ταξινομητή Bernulli, εξετάστηκαν διαφορετικές τιμές των *εκ των προτέρων πιθανοτήτων* (prior probabilities) των κλάσεων. Οι πιθανότητες αυτές δηλώνουν την κατανομή πιθανότητας των κλάσεων που θα μπορούσε κάποιος να υποθέσει πριν την εκπαίδευση του ταξινομητή και εξαρτάται κυρίως από τα χαρακτηριστικά του προβλήματος. Στην περίπτωση του προβλήματος της πρόβλεψης μελλοντικών ακμών στα μέσα κοινωνικής δικτύωσης, όπως έχει προαναφερθεί, αναμένεται ο αριθμός των δειγμάτων της αρνητικής κλάσης να είναι υπερβολικά μεγαλύτερος από τον αριθμό των δειγμάτων της θετικής κλάσης. Η βέλτιστη τιμή αυτής της παραμέτρου ορίστηκε στο $0,8$ για την κλάση 0 και $0,2$ για την κλάση 1.

Επίσης, η χρήση του ταξινομητή κατανομής Bernulli προϋποθέτει την ύπαρξη χαρακτηριστικών εισόδου με δυαδικές τιμές (αληθές / ψευδές). Τα χαρακτηριστικά που συνθέτουν το διάνυσμα εισόδου



(a) Συλλογή δεδομένων `soc-hamsterster`



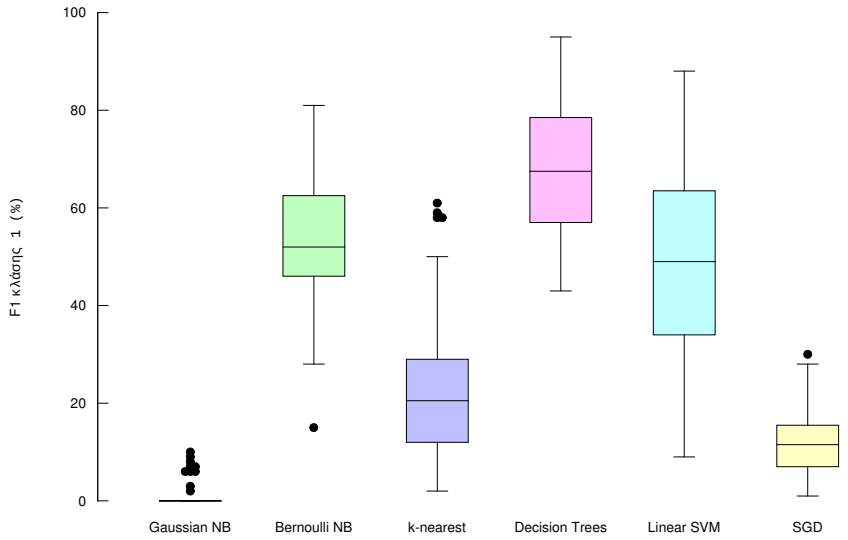
(b) Συλλογή δεδομένων `soc-advogato`

Σχήμα 5.6: Ανάκληση ευφυών τεχνικών για την θετική κλάση (διασταυρούμενη αντεπικύρωση 100 διπλωμάτων)

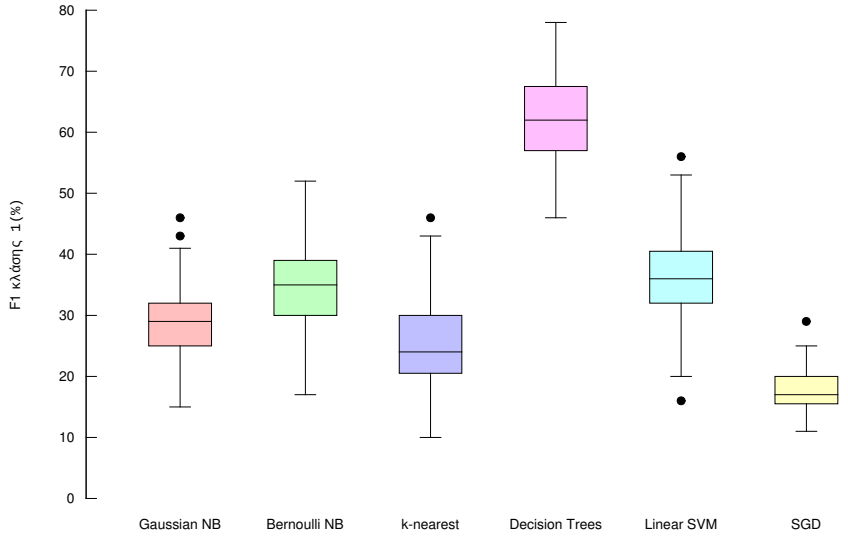
στην περίπτωση που εξετάζεται, λαμβάνουν συνεχείς πραγματικές τιμές και άρα θα πρέπει να χρησιμοποιηθεί κάποιο κατώφλι για την μετατροπή των πραγματικών τιμών σε δυαδικές. Αφού εξετάστηκε η απόδοση του ταξινομητή για διαφορετικές τιμές κατωφλίου και στα δυο σύνολα δεδομένων, βρέθηκε ότι η βέλτιστη τιμή για το εν λόγω κατώφλι είναι το $1,8$.

Στον ταξινομητή k-πλησιέστερων γειτόνων, μελετήθηκε η απόδοση για διαφορετικό αριθμό πλησιέστερων γειτόνων (k) που πρέπει να ελεγχθούν ως προς την κλάση στην οποία ανήκουν, καθώς και για διαφορετικό τρόπο υπολογισμού της απόστασης μεταξύ των δειγμάτων. Συγκεκριμένα, μελετήθηκε η Ευκλείδεια απόσταση και η απόσταση Manhattan (Ενότητα 3.3.2). Ο βέλτιστος αριθμός γειτόνων βρέθηκε να είναι το 2, ενώ η απόσταση Manhattan εμφάνισε οριακά καλύτερα αποτελέσματα. Περαιτέρω ανάλυση των παραμέτρων που επηρεάζουν την απόδοση του ταξινομητή παρουσιάζεται στην Ενότητα 5.2.3..

Στα δέντρα αποφάσεων, εξετάστηκαν διαφορετικές συναρτήσεις υπολογισμού της ποιότητας ενός διαχωρισμού και πιο συγκεκριμένα ο δείκτης Gini και η πληροφοριακή απολαβή (Ενότητα 3.4.2). Επί-



(a) Συλλογή δεδομένων `soc-hamsterster`



(b) Συλλογή δεδομένων `soc-advogato`

Σχήμα 5.7: Μετρική F1 ευφυών τεχνικών για την θετική κλάση (διασταυρούμενη αντεπικύρωση 100 διπλωμάτων)

σης, εξετάστηκαν διαφορετικές στρατηγικές διαχωρισμού για διαφορετικό πλήθος χαρακτηριστικών που λήφθηκαν υπόψη κατά την διαδικασία του διαχωρισμού (όλα τα χαρακτηριστικά, λογάριθμος του πλήθους των χαρακτηριστικών, τετραγωνική ρίζα του πλήθους των χαρακτηριστικών). Τέλος, μελετήθηκε η απόδοση των δέντρων αποφάσεων για ισοδύναμα και ισορροπημένα βάρη των κλάσεων. Η τιμή των ισορροπημένων βαρών των κλάσεων επηρεάζεται αντιστρόφως ανάλογα από την συχνότητα εμφάνισης της ετικέτας τους στα δεδομένα εισόδου. Η βέλτιστη απόδοση βρέθηκε για χρήση του δείκτη Gini σαν συνάρτηση υπολογισμού της ποιότητας διαχωρισμού, χρήση όλων των χαρακτηριστικών κατά την διαδικασία διαχωρισμού και για ισορροπημένα βάρη στις δυο κλάσεις του προβλήματος.

Στις μηχανές διανυσμάτων υποστήριξης εξετάστηκαν διαφορετικές τιμές της παραμέτρου C που καθορίζει την επιρροή του σφάλματος στη διαδικασία μάθησης, για διαφορετικές συναρτήσεις σφάλματος (hinge, τετραγωνικό hinge). Κατά την διαδικασία υπολογισμού της συνάρτησης κόστους ερευνήθηκε η απόδοση με χρήση των L_1 και L_2 νορμών για τον υπολογισμό του όρου $||w||$, αλλά και χρήση διαφορετικών βαρών στις κλάσεις για τον υπολογισμό του όρου $jC \sum_{i,y_i=1} \xi_i + C \sum_{i,y_i=-1} \xi_i$

της Εξίσωσης 3.21 (Ενότητα 3.5.3). Η βέλτιστη απόδοση βρέθηκε για τιμή της παραμέτρου C το 1, χρήση της L_1 νόρμας για τον υπολογισμού του όρου $||w||$ και χρήση ισορροπημένων βαρών στις κλάσεις. Δεν βρέθηκε σημαντική αλλαγή στην απόδοση του ταξινομητή για τις διαφορετικές συναρτήσεις σφάλματος.

Τέλος, και στην μελέτη της στοχαστικής καθόδου κλίσης, εξετάστηκαν διαφορετικές συναρτήσεις σφάλματος (hinge, λογαριθμική απώλεια, τροποποιημένη huber, τετραγωνική hinge, γραμμική απώλεια), διαφορετικοί τρόποι υπολογισμού του όρου κανονικοποίησης R και πιο συγκεκριμένα οι L_1 και L_2 νόρμες (Ενότητα 3.6.1) καθώς και διαφορετικά βάρη στις κλάσεις. Βέλτιστη απόδοση επιτεύχθηκε για τη λογαριθμική συνάρτηση σφάλματος, χρήση της L_2 νόρμας για τον υπολογισμό του όρου κανονικοποίησης και ισορροπημένα βάρη στις κλάσεις.

5.2.2 Αποτελέσματα

Όπως και στις γραφοθεωρητικές τεχνικές, παρατηρείται ότι η διασπορά των διαφόρων ευφυών τεχνικών είναι διαφορετική για τα δυο σύνολα δεδομένων. Συγκεκριμένα, στο σύνολο `soc-hamsterster` παρατηρείται μεγαλύτερη απόκλιση από την μέση τιμή για όλους τους ταξινομητές που μελετήθηκαν (Σχήματα 5.5, 5.6 και 5.7). Όπως προαναφέρθηκε, το σύνολο δεδομένων `soc-advogato` είναι μεγαλύτερο με αποτέλεσμα σε κάθε επανάληψη της μεθόδου διασταυρούμενης αντεπικύρωσης 100 διπλωμάτων το σύνολο ελέγχου να περιέχει περισσότερα στον αριθμό ζεύγη κόμβων, από την αντίστοιχη περίπτωση για το σύνολο `soc-hamsterster`. Η λανθασμένη ταξινόμηση ενός ζεύγους κόμβων από το σύνολο ελέγχου στο σύνολο δεδομένων `soc-advogato` επιφέρει ποινή χαμηλότερης βαρύτητας στον μέσο όρο της απόδοσης του εκάστοτε ταξινομητή για την συγκεκριμένη επανάληψη της μεθόδου διασταυρούμενης αντεπικύρωσης και άρα και στην συνολική απόδοσή του.

Η συμπεριφορά των ταξινομητών που μελετούνται φαίνεται να είναι παρόμοια στα δυο σύνολα δεδομένων, με μόνη διαφορά τον γκαουσιανό απλό μπεϋζιανό ταξινομητή, όπου η απόδοση του είναι πολύ χαμηλή για το σύνολο δεδομένων `soc-hamsterster` και μέτρια για το σύνολο δεδομένων `soc-advogato`. Η χαμηλή αυτή απόδοση οφείλεται στο γεγονός πως τα στοιχεία του διανύσματος χαρακτηριστικών της εισόδου του ταξινομητή (Ενότητα 4.4) απέχουν πολύ από το να ακολουθούν την κανονική κατανομή. Κατά συνέπεια, ο εν λόγω ταξινομητής αδυνατεί να «μάθει» από τα δεδομένα εκπαίδευσης και κατόπιν να γενικεύσει. Είναι δε χαρακτηριστικό ότι ο γκαουσιανός απλός μπεϋζιανός ταξινομητής όχι μόνο εμφανίζει τη χειρότερη απόδοση μεταξύ όλων των ευφυών τεχνικών στο `soc-hamsterster` αλλά υπολείπεται ακόμα και των γραφοθεωρητικών τεχνικών.

Σε αντίθεση με την προαναφερόμενη περίπτωση, η διακριτοποίηση στη δυαδική κλίμακα του διανύσματος χαρακτηριστικών του απλού μπεϋζιανού ταξινομητή κατανομής Bernulli (Ενότητα 5.2.1) του επιτρέπει να ανακαλύπτει καλύτερα τις σχέσεις που διέπουν τα δεδομένα εκπαίδευσης και συνεπώς να ταξινομεί καλύτερα τα δείγματα του εκάστοτε συνόλου ελέγχου. Χαρακτηριστικό του συγκεκριμένου ταξινομητή είναι ότι επιτυγχάνει και μια καλή ισορροπία μεταξύ Ακρίβειας και Ανάκλησης, η οποία αποτυπώνεται και στην συνδυαστική μετρική F1. Είναι επίσης αξιοσημείωτο ότι ο απλός μπεϋζιανός ταξινομητής Bernoulli επιτυγχάνει την δεύτερη καλύτερη επίδοση και στις δύο συλλογές δεδομένων που εξετάστηκαν.

Από την άλλη, ο ταξινομητής των k-πλησιέστερων γειτόνων φαίνεται να έχει μεγάλη απόκλιση από την μέση τιμή και στις τρεις μετρικές απόδοσης. Αυτό σημαίνει ότι σε ορισμένες επαναλήψεις της μεθόδου διασταυρούμενης αντεπικύρωσης, ο ταξινομητής πετυχαίνει πολύ καλή απόδοση στα συγκεκριμένα σύνολα ελέγχου, ενώ σε κάποιες άλλες περιπτώσεις η απόδοση του μειώνεται δραματικά. Αυτή η συμπεριφορά αποδίδεται στο γεγονός πως η μικρή απόσταση μεταξύ δυο ή περισσοτέρων διανυσμάτων χαρακτηριστικών (δειγμάτων) δεν εξασφαλίζει πάντα την ταξινόμηση τους στην ίδια κλάση. Ας θεωρήσουμε το απλό παράδειγμα δυο δειγμάτων που αναπαρίστανται με διανύσματα τριών χαρακτηριστικών. Το διάνυσμα αποτελούν ο βαθμός των δυο κόμβων και η μετρική των κοινών γειτόνων. Ας υποθέσουμε ότι τα δυο δείγματα έχουν διανύσματα τα $[6, 6, 3]$ και $[6, 7, 3]$. Η απόσταση μεταξύ των δυο διανυσμάτων είναι πολύ μικρή, αφού και οι τέσσερις κόμβοι έχουν σχεδόν τον ίδιο βαθμό ενώ τα ζεύγη των κόμβων έχουν τον ίδιο αριθμό κοινών γειτόνων. Παρόλα αυτά, η μικρή απόσταση των διανυσμάτων δεν εγγυάται ότι τα δυο δείγματα ανήκουν στην ίδια κλάση.

Το προαναφερόμενο παράδειγμα εξηγεί και το λόγο που η απόδοση του ταξινομητή των k-πλησιέστερων γειτόνων είναι καλύτερη στο πιο αραιό σύνολο δεδομένων (`soc-advogato`), μιας και εκεί η απόσταση μεταξύ των δειγμάτων που ανήκουν σε διαφορετικές κλάσεις είναι μεγαλύτερη. Συνολικότερα, ωστόσο, ο συγκεκριμένος ταξινομητής επιτυγχάνει από τις χειρότερες αποδόσεις μεταξύ των ευφυών τεχνικών, πράγμα το οποίο σημαίνει ότι αδυνατεί να λειτουργήσει ικανοποιητικά σε περιπτώσεις συνόλων δεδομένων με έντονο το φαινόμενο της ανισορροπίας των κλάσεων, όπως αυτή που εξετάζεται στην παρούσα διπλωματική εργασία.

Εντελώς αντίθετη είναι η εικόνα όσον αφορά τα δέντρα αποφάσεων, που εμφανίζουν με διαφορά την καλύτερη απόδοση και στα δύο σύνολα δεδομένων και για τις τρεις μετρικές. Εκτός αυτών, επιτυγχάνουν τη ζητούμενη ισορροπία μεταξύ Ακρίβειας και Ανάκλησης ενώ επιπρόσθετα παρουσιάζουν και τη μικρότερη διασπορά τιμών. Είναι δε χαρακτηριστικό ότι και στο `soc-hamsterster` και στο `soc-advogato` έχουν απόδοση μεγαλύτερη του 60% για την ταξινόμηση των δειγμάτων της θετική κλάσης, ποσοστό αρκετά υψηλό, ειδικά αν ληφθεί υπόψη ο μεγάλος βαθμός ανισορροπίας των κλάσεων που παρουσιάζει το πρόβλημα που εξετάζεται. Συνεπώς, ένα γενικό συμπέρασμα της παρούσας διατριβής είναι πως η χρήση των δέντρων αποφάσεων είναι η πλέον ενδεδειγμένη λύση για το πρόβλημα της πρόβλεψης μελλοντικών ακμών μεταξύ των χρηστών των μέσων κοινωνικής δικτύωσης.

Οι μηχανές διανυσμάτων υποστήριξης με γραμμικό πυρήνα είναι η τεχνική που εμφάνισε τη μεγαλύτερη διασπορά τιμών (από όσες εξετάστηκαν), ειδικά στη συλλογή δεδομένων `soc-hamsterster`. Αυτή η συμπεριφορά είναι πιθανό να οφείλεται στο γεγονός πως τα δεδομένα εισόδου ήταν μικρών διαστάσεων (διανύσματα 8 χαρακτηριστικών) ενώ είναι γνωστό πως η χρήση της εν λόγω τεχνικής ενδεικνύεται για δεδομένα μεγάλων διαστάσεων. Ένας ακόμα πιθανός λόγος είναι και η γραμμική συνάρτηση πυρήνα που χρησιμοποιήθηκε. Ενδεχομένως αν είχε χρησιμοποιηθεί διαφορετική συνάρτηση απόφασης, που να ανταποκρίνεται καλύτερα στα χαρακτηριστικά του προβλήματος, η διασπορά των τιμών να ήταν μικρότερη και η συνολική εικόνα του ταξινομητή πολύ καλύτερη.

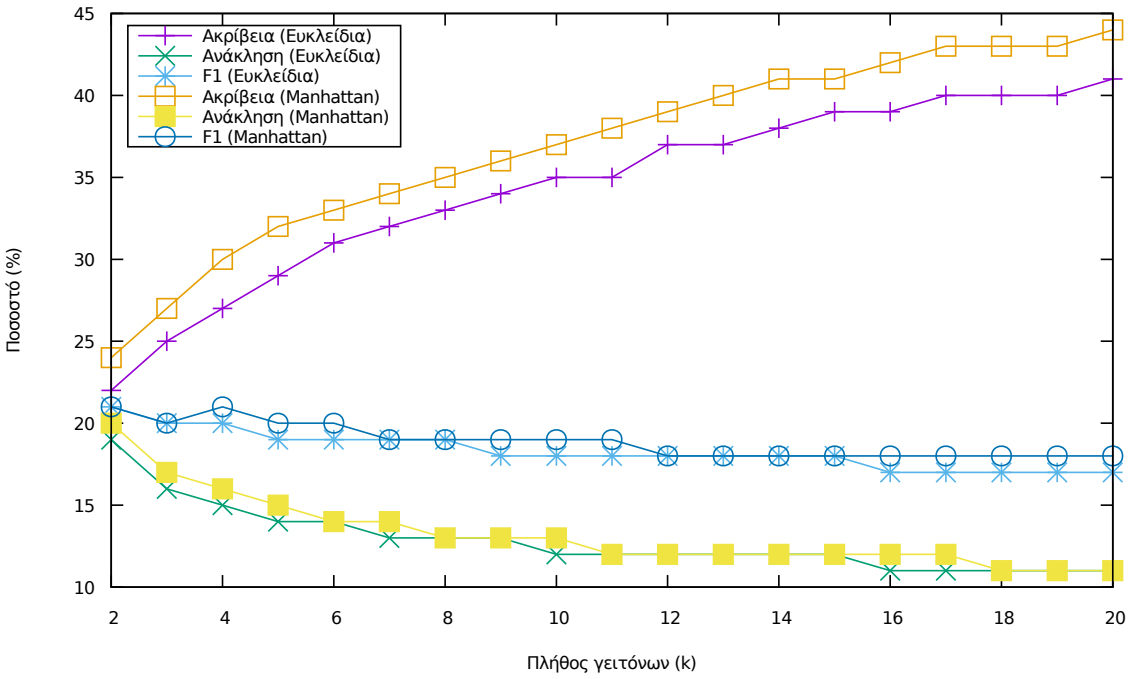
Τέλος, η μέθοδος της στοχαστικής καθόδου κλίσης εμφανίζει υψηλό βαθμό Ανάκλησης και πολύ χαμηλό βαθμό Ακρίβειας και στις δύο συλλογές δεδομένων. Με άλλα λόγια, ενώ είναι σε θέση να εντοπίζει την πλειοψηφία των δειγμάτων της θετικής κλάσης στο σύνολο ελέγχου, ωστόσο κάνει και πολλές προτάσεις που είναι λανθασμένες, με αποτέλεσμα να εμφανίζει το χειρότερο αποτέλεσμα μεταξύ των ευφυών τεχνικών, όσον αφορά τη μετρική F1 (με την προφανή εξαίρεση του απλού μπεϋζιανού γκαουσιανού ταξινομητή στο `soc-hamsterster`). Και σε αυτή την περίπτωση, παρόλα αυτά, το αποτέλεσμα δικαιολογείται από τα χαρακτηριστικά του προβλήματος που εξετάζεται, μιας και η συγκεκριμένη τεχνική αποδίδει καλύτερα όταν τόσο τα δεδομένα όσο και η διάσταση του διανύσματος των χαρακτηριστικών είναι μεγάλες, το οποίο όμως δεν ισχύει στο πρόβλημα που εξετάζεται.

5.2.3 Μελέτη ταξινομητή k-πλησιέστερων γειτόνων

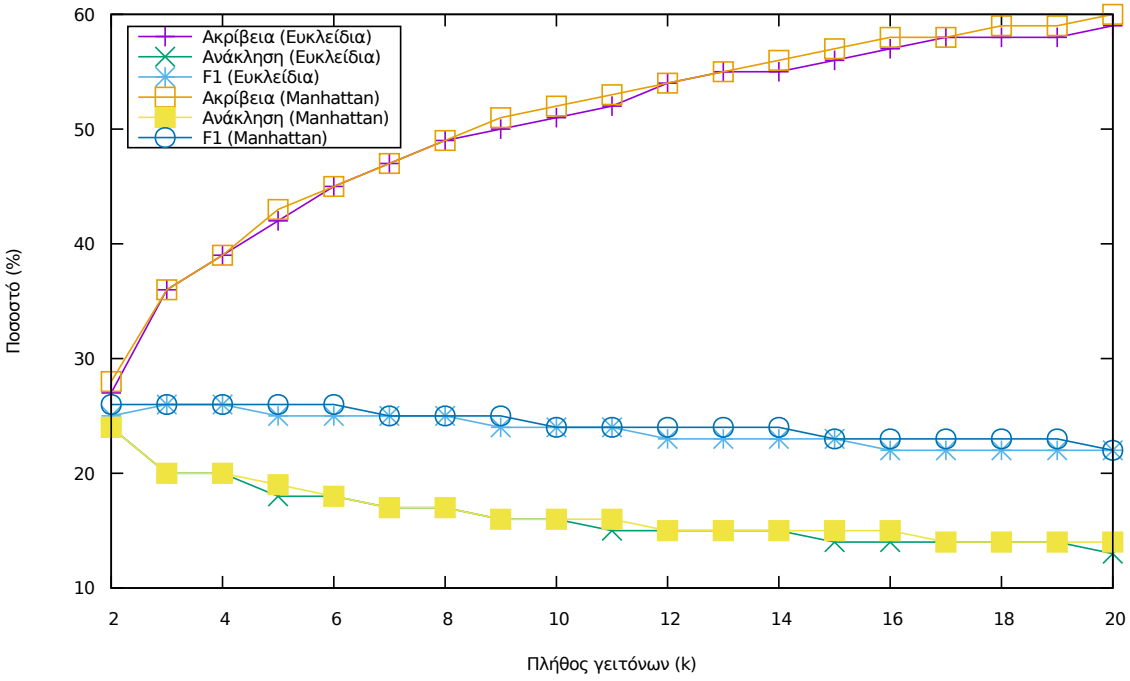
Σε αυτήν την ενότητα θα μελετηθεί λεπτομερώς η απόδοση του ταξινομητή k-πλησιέστερων γειτόνων ως προς τον αριθμό των γειτόνων. Όπως ειπώθηκε στην Ενότητα 3.3.2, η ταξινόμηση ενός νέου δείγματος σε μια κλάση του προβλήματος επηρεάζεται σαφώς από την κλάση των k κοντινότερων δειγμάτων (Σχήμα 3.5).

Στο Σχήμα 5.8 παρουσιάζονται οι τρεις μετρικές απόδοσης για τα δυο σύνολα δεδομένων με χρήση του ταξινομητή k-πλησιέστερων γειτόνων. Στο ίδιο σχήμα περιλαμβάνονται και οι δυο τρόποι υπολογισμού της απόστασης μεταξύ των γειτόνων (Ευκλείδεια απόσταση και απόσταση Manhattan), παρότι η συμπεριφορά του ταξινομητή είναι σχεδόν ίδια. Με την αύξηση του αριθμού των γειτόνων παρατηρείται σημαντική αύξηση της ακρίβειας και πτώση της ανάκλησης και στα δυο σύνολα δεδομένων.

Η εξήγηση είναι απλή και βασίζεται στα χαρακτηριστικά του προβλήματος και πιο συγκεκριμένα σε αυτό της ανισορροπίας των κλάσεων. Είναι γνωστό πως το συντριπτικό ποσοστό ζευγών κόμβων του συνόλου εκπαίδευσης ανήκει στην αρνητική κλάση του προβλήματος. Κατά την διαδικασία ταξινόμησης νέων δειγμάτων, ο έλεγχος μεγαλύτερου αριθμού πλησιέστερων γειτόνων έχει ως αποτέλεσμα την κατηγοριοποίηση μειωμένου αριθμού δειγμάτων στην θετική κλάση, εφόσον στις περισσότερες περιπτώσεις ο αριθμός των γειτονικών δειγμάτων αρνητικής κλάσης θα είναι μεγαλύτερος.



(a) Συλλογή δεδομένων `soc-hamsterster`



(b) Συλλογή δεδομένων `soc-advogato`

Σχήμα 5.8: Απόδοση ταξινομητή k-πλησιέστερων γειτόνων ως προς τον αριθμό γειτόνων (διασταυρούμενη αντεπικύρωση 100 διπλωμάτων)

Αυτή η μείωση στον αριθμό προβλέψεων ζευγών κόμβων της θετικής κλάσης, επιφέρει μείωση στην Ανάκληση αφού πλέον ο ταξινομητής έχει μικρότερη δυνατότητα να προβλέψει όλα τα θετικά δείγματα του συνόλου ελέγχου. Αντίθετα, η Ακρίβεια αυξάνεται εφόσον ο ταξινομητής κάνει λιγότερες, πιο σίγουρες και πιο σωστές προβλέψεις για την θετική κλάση.

Κεφάλαιο 6

Συμπεράσματα και μελλοντικές κατευθύνσεις

6.1 Συμπεράσματα

Σε αυτήν τη διπλωματική εργασία μελετήθηκε το πρόβλημα της πρόβλεψης μελλοντικών ακμών μεταξύ των χρηστών των μέσων κοινωνικής δικτύωσης και παρουσιάστηκαν διάφορες τεχνικές επίλυσης. Συγκεκριμένα, μελετήθηκαν γραφοθεωρητικές τεχνικές που εξετάζουν την δομή του δικτύου, με σκοπό τον υπολογισμό ενός βαθμού ομοιότητας μεταξύ των κόμβων. Επίσης, μελετήθηκε η απόδοση διαφόρων ταξινομητών, αφού πρώτα ορίστηκε το αντίστοιχο πρόβλημα δυαδικής κατηγοριοποίησης.

Όπως αναφέρθηκε σε πολλά σημεία της παρούσας εργασίας, τα μέσα κοινωνικής δικτύωσης παρουσιάζουν χαρακτηριστικά δικτύων ελεύθερης κλίμακας και η αραιότητα είναι βασική ιδιότητά τους. Εξαιτίας αυτών των χαρακτηριστικών η μελέτη και εφαρμογή των τεχνικών επίλυσης του προβλήματος πρόβλεψης ακμών, γίνεται ιδιαίτερα δύσκολη και απαιτητική. Συγκεκριμένα, τεχνικές που βασίζονται στην μελέτη της δομής του δικτύου παρουσιάζουν χαμηλή απόδοση λόγω της προβληματικής δομής του γράφου του δικτύου, ενώ το πρόβλημα της ανισορροπίας των κλάσεων δυσκολεύει την πρόβλεψη με χρήση ευφυών τεχνικών.

Παρόλα αυτά, όπως φάνηκε από τα πειραματικά αποτελέσματα της εργασίας (Κεφάλαιο 5), ακόμα και η γενική παρατήρηση της δομής του γράφου μπορεί να επιφέρει καλύτερα αποτελέσματα σε σχέση με την τυχαία πρόβλεψη μελλοντικών ακμών. Αν και η απόδοση των γραφοθεωρητικών τεχνικών βρέθηκε να είναι αρκετά χαμηλή, οι τρεις τεχνικές που φάνηκαν να υπερισχύουν έναντι των άλλων είναι η μετρική των κοινών γειτόνων, η μετρική Jaccard και η μετρική Adamic-Adar, των οποίων ο ορισμός είναι σχετικά όμοιος. Πρέπει επίσης να τονιστεί η πάρα πολύ χαμηλή απόδοση των τεχνικών που στηρίζονται στα μονοπάτια και στους τυχαίους περιπάτους, όπου η αραιότητα των μέσων κοινωνικής δικτύωσης φαίνεται να διαδραμάτισε καθοριστικό ρόλο στην δραματική μείωση της απόδοσής τους.

Το σημαντικότερο πρόβλημα που αντιμετωπίσαμε στις ευφυείς τεχνικές ήταν αυτό της ανισορροπίας των δυο κλάσεων και συγκεκριμένα ο υπερβολικά μικρός αριθμός θετικών δειγμάτων στο σύνολο εκπαίδευσης. Οι ρίζες του προβλήματος αυτού βρίσκονται στην αραιότητα των μέσων κοινωνικής δικτύωσης και επηρεάζουν την απόδοση πολλών εκ των ταξινομητών που μελετήσαμε. Η απόδοση των ευφυών τεχνικών, σύμφωνα με τα αποτελέσματά μας, ήταν πολύ καλύτερη των γραφοθεωρητικών τεχνικών στις περισσότερες περιπτώσεις. Τα δέντρα αποφάσεων φαίνονται να έχουν την καλύτερη απόδοση, ενώ πρέπει να σημειωθεί η προβληματική εικόνα που εμφάνισαν ο ταξινομητής k-πλησιέστερων γειτόνων, οι μηχανές διανυσμάτων υποστήριξης και ο αλγόριθμος στοχαστικής καθόδου κλίσης.

6.2 Μελλοντικές κατευθύνσεις

Το πρόβλημα της πρόβλεψης ακμών μεταξύ των χρηστών των κοινωνικών δικτύων παραμένει ένα αρκετά ενεργό επιστημονικό αντικείμενο, το οποίο δεν περιορίζεται μόνο στις τεχνικές που παρουσιάστηκαν στην παρούσα εργασία. Σε μια πρώτη φάση θα ήταν χρήσιμο να μελετηθεί η μεταβολή της απόδοσης των απλών τεχνικών σε σχέση με διάφορα χαρακτηριστικά του δικτύου. Για παράδειγμα, εφόσον είναι γνωστό ότι η αραιότητα των κοινωνικών δικτύων επηρεάζει αρνητικά την απόδοση των απλών τεχνικών, θα μπορούσε να εξεταστεί η μεταβολή στην απόδοση μιας τεχνικής σε σχέση με μια

πιθανή μεταβολή στον μέσο βαθμό των κόμβων ή την πυκνότητα του δικτύου.

Επίσης, στην παρούσα εργασία μελετήθηκαν στατικά στιγμιότυπα του δικτύου, για μια συγκεκριμένη χρονική στιγμή. Παρόλα αυτά, τα μέσα κοινωνικής δικτύωσης είναι δυναμικές οντότητες που εξελίσσονται και παρακμάζουν με την πάροδο του χρόνου, με νέους κόμβους και ακμές να δημιουργούνται και να καταστρέφονται κάθε στιγμή. Θα ήταν επομένως χρήσιμη η δυναμική μελέτη τέτοιων δικτύων, χρησιμοποιώντας χρονικές ετικέτες για τους κόμβους και τις ακμές. Για παράδειγμα, σε ένα μέσο κοινωνικής δικτύωσης σαν το Facebook, μη ενεργοί χρήστες, δηλαδή κόμβοι του δικτύου που η χρονική ετικέτα τους δηλώνει ότι η τελευταία δραστηριότητά τους ήταν πριν πολλούς μήνες, θα μπορούσαν να διαγραφούν απλοποιώντας έτσι την δομή του δικτύου. Επιπλέον, οι πιο πρόσφατα δημιουργημένες ακμές μεταξύ κόμβων θα μπορούσαν να έχουν μεγαλύτερο συντελεστή βαρύτητας στον υπολογισμό της απόδοσης των γραφοθεωρητικών τεχνικών.

Εκτός από την παράμετρο του χρόνου, διάφορα χαρακτηριστικά των χρηστών θα μπορούσαν να χρησιμοποιηθούν για τον υπολογισμό του βαθμού ομοιότητας μεταξύ τους, εφόσον είναι διαθέσιμα. Πληροφορίες όπως η ηλικία, η περιοχή κατοικίας, η εκπαίδευση, τα χόμπι, αλλά και το είδος μουσικής και ταινιών που αρέσει σε έναν χρήστη θα μπορούσαν να αποτελέσουν μέτρο ομοιότητας μεταξύ χρηστών του δικτύου.

Όσον αφορά τις ευφυείς τεχνικές μια σημαντική παράμετρος είναι η είσοδος των ταξινομητών, δηλαδή το διάνυσμα χαρακτηριστικών των δειγμάτων εκπαίδευσης. Θα ήταν χρήσιμο να μελετηθεί η απόδοση των ταξινομητών σε σχέση με τον αριθμό των χαρακτηριστικών που χρησιμοποιούνται, καθώς και τους διάφορους συνδυασμούς μετρικών που μπορούν να αποτελέσουν το διάνυσμα. Η χρήση διαφορετικών γραφοθεωρητικών τεχνικών ως στοιχείων του διανύσματος χαρακτηριστικών αλλά και η χρήση άλλων χαρακτηριστικών του ζεύγους κόμβων που αποτελεί το δείγμα εκπαίδευσης, όπως για παράδειγμα η ηλικία των χρηστών, θα μπορούσαν να επιφέρουν καλύτερη απόδοση.

Επίσης, θα μπορούσαν να εξεταστούν διαφορετικές τεχνικές επίλυσης του προβλήματος της ανισορροπίας των δυο κλάσεων. Πιο συγκεκριμένα, η δημιουργία τυχαίων θετικών δειγμάτων με σκοπό την ενίσχυση της θετικής κλάσης (oversampling) και την εξισορρόπηση των ποσοστών των δειγμάτων των δυο κλάσεων στο σύνολο εκπαίδευσης, θα μπορούσε να αποτελέσει μια εναλλακτική λύση του προβλήματος.

Τέλος, η παρούσα διπλωματική εργασία επικεντρώθηκε σε ορισμένους ταξινομητές όπως ο μπεϋζιανός ταξινομητής, ο ταξινομητής των k-πλησιέστερων γειτόνων, τα δέντρα αποφάσεων, οι μηχανές διανυσμάτων υποστήριξης και ο ταξινομητής της μεθόδου στοχαστικής καθόδου κλίσης. Το πρόβλημα της πρόβλεψης μελλοντικών ακμών μεταξύ των χρηστών των μέσων κοινωνικής δικτύωσης θα μπορούσε να μελετηθεί και με χρήση πιο περίπλοκων ευφυών τεχνικών, όπως τα *τυχαία δάση* (random forests), η μέθοδος της *προσομοιωμένης ανόπτησης* (simulated annealing), τα *νευρωνικά δίκτυα* (neural networks), καθώς και σύνθετων συνδυασμών τους.

Βιβλιογραφία

[Adaf05] Sisay Fissaha Adafre and Maarten de Rijke, "Discovering Missing Links in Wikipedia", in *Proceedings of the 3rd International Workshop on Link Discovery*, LinkKDD '05, pp. 90–97, New York, NY, USA, 2005, ACM.

[Adam01] Lada Adamic and Eytan Adar, "Friends and Neighbors on the Web", *Social Networks*, vol. 25, pp. 211–230, 2001.

[Airo08] Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg and Eric P. Xing, "Mixed Membership Stochastic Blockmodels", *J. Mach. Learn. Res.*, vol. 9, pp. 1981–2014, June 2008.

[Ande12] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg and Jure Leskovec, "Effects of User Similarity in Social Media", in *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pp. 703–712, New York, NY, USA, 2012, ACM.

[Bara02] A. L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert and T. Vicsek, "Evolution of the social network of scientific collaborations", 2002.

[Bhat11] Prantik Bhattacharyya, Ankush Garg and Shyhtsun Felix Wu, "Analysis of user keyword similarity in online social networks", *Social Network Analysis and Mining*, vol. 1, no. 3, pp. 143–158, 2011.

[Bigg93] Norman Biggs, *Algebraic Graph Theory*, Cambridge University Press, Cambridge, 2nd edition, 1993.

[Bose92] Bernhard E. Boser, Isabelle M. Guyon and Vladimir N. Vapnik, "A Training Algorithm for Optimal Margin Classifiers", in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pp. 144–152, New York, NY, USA, 1992, ACM.

[Bous08] Olivier Bousquet and Léon Bottou, "The Tradeoffs of Large Scale Learning", in J. C. Platt, D. Koller, Y. Singer and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pp. 161–168, Curran Associates, Inc., 2008.

[Bran03] Janez Brank, Marko Grobelnik, Natasa Milic-Frayling and Dunja Mladenic, "Training text classifiers with SVM on very few positive examples", Technical Report MSR-TR-2003-34, Microsoft Research, April 2003.

[Brei84] Leo Breiman et al., *Classification and Regression Trees*, Chapman & Hall, New York, 1984.

[Brem05] David Bremner, Erik Demaine, Jeff Erickson, John Iacono, Stefan Langerman, Pat Morin and Godfried Toussaint, "Output-Sensitive Algorithms for Computing Nearest-Neighbour Decision Boundaries", *Discrete & Computational Geometry*, vol. 33, no. 4, pp. 593–604, 2005.

[Chen12] Hung-Hsuan Chen, Liang Gou, Xiaolong (Luke) Zhang and C. Lee Giles, "Discovering Missing Links in Networks Using Vertex Similarity Measures", in *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, SAC '12, pp. 138–143, New York, NY, USA, 2012, ACM.

[Coom82] D. Coomans and D.L. Massart, "Alternative k-nearest neighbour rules in supervised pattern recognition", *Analytica Chimica Acta*, vol. 136, pp. 15 – 27, 1982.

[Cort95] Corinna Cortes and Vladimir Vapnik, "Support-vector networks", *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[Fous07] Francois Fouss, Alain Pirotte, Jean-Michel Renders and Marco Saerens, "Random-Walk Computation of Similarities Between Nodes of a Graph with Application to Collaborative Recommendation", *IEEE Trans. on Knowl. and Data Eng.*, vol. 19, no. 3, pp. 355–369, March 2007.

[Gala85] Galaskiewicz, "Copyright", in Joseph Galaskiewicz, editor, *Social Organization of an Urban Grants Economy*, pp. iv –, Academic Press, 1985.

[Hast83] Reid. Hastie, Steven. Penrod and Nancy. Pennington, *Inside the jury / Reid Hastie, Steven D. Penrod, Nancy Pennington*, Harvard University Press Cambridge, Mass, 1983.

[Hoth06] Torsten Hothorn, Kurt Hornik and Achim Zeileis, "Unbiased recursive partitioning: A conditional inference framework", *JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS*, vol. 15, no. 3, pp. 651–674, 2006.

[Huan05] Zan Huang, Xin Li and Hsinchun Chen, "Link Prediction Approach to Collaborative Filtering", in *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '05, pp. 141–142, New York, NY, USA, 2005, ACM.

[Hyaf76] Laurent Hyafil and Ronald L. Rivest, "Constructing Optimal Binary Decision Trees is NP-Complete.", *Inf. Process. Lett.*, vol. 5, no. 1, pp. 15–17, 1976.

[Jeh02] Glen Jeh and Jennifer Widom, "SimRank: A Measure of Structural-context Similarity", in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pp. 538–543, New York, NY, USA, 2002, ACM.

[Katz53] Leo Katz, "A new status index derived from sociometric analysis", *Psychometrika*, vol. 18, no. 1, pp. 39–43, March 1953.

[Leic06] E. A. Leicht, P. Holme and M. E. J. Newman, "Vertex similarity in networks", *Physical Review E*, vol. 73, no. 2, p. 026120, February 2006.

[Levi99] Raph Levien, "Advogato Online Community", *http://www.advogato.org/*, 1999.

[Li09] Xin Li and Hsinchun Chen, "Recommendation As Link Prediction: A Graph Kernel-based Machine Learning Approach", in *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '09, pp. 213–216, New York, NY, USA, 2009, ACM.

[Libe03] David Liben-Nowell and Jon Kleinberg, "The Link Prediction Problem for Social Networks", in *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, pp. 556–559, New York, NY, USA, 2003, ACM.

[Lich10] Ryan N. Lichtenwalter, Jake T. Lussier and Nitesh V. Chawla, "New Perspectives and Methods in Link Prediction", in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pp. 243–252, New York, NY, USA, 2010, ACM.

[Liu13] Haifeng Liu, Zheng Hu, Hamed Haddadi and Hui Tian, "Hidden link prediction based on node centrality and weak ties", *EPL (Europhysics Letters)*, pp. 18004+, January 2013.

[Lova96] L. Lovász, "Random Walks on Graphs: A Survey", in D. Miklós, V. T. Sós and T. Szőnyi, editors, *Combinatorics, Paul Erdős is Eighty*, vol. 2, pp. 353–398, János Bolyai Mathematical Society, Budapest, 1996.

[Lu09] Linyuan Lü, Ci-Hang Jin and Tao Zhou, "Similarity index based on local paths for link prediction of complex networks", *Phys. Rev. E*, vol. 80, p. 046122, Oct 2009.

[McCa98] Andrew McCallum and Kamal Nigam, "A comparison of event models for Naive Bayes text classification", 1998.

[Newm01] M.E.J. Newman, "Clustering and preferential attachment in growing networks", *Physical Review E*, vol. 64, no. 2, p. 025102, 2001.

[Papa12] Alexis Papadimitriou, Panagiotis Symeonidis and Yannis Manolopoulos, "Fast and Accurate Link Prediction in Social Networking Systems", *J. Syst. Softw.*, vol. 85, no. 9, pp. 2119–2132, September 2012.

[Quin86] J. R. Quinlan, "Induction of Decision Trees", *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, March 1986.

[Rava02] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai and A.-L. Barabási, "Hierarchical Organization of Modularity in Metabolic Networks", *Science*, vol. 297, no. 5586, pp. 1551–1555, 2002.

[Roka05] L. Rokach and O. Maimon, "Top-down Induction of Decision Trees Classifiers - a Survey", *Trans. Sys. Man Cyber Part C*, vol. 35, no. 4, pp. 476–487, November 2005.

[Roka08] Lior Rokach and Oded Maimon, *Data Mining with Decision Trees: Theroy and Applications*, World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2008.

[Ross13a] Ryan A. Rossi and Nesreen K. Ahmed, "soc-advogato - Social Networks", 2013.

[Ross13b] Ryan A. Rossi and Nesreen K. Ahmed, "soc-hamsterster - Social Networks", 2013.

[Ross15] Ryan A. Rossi and Nesreen K. Ahmed, "The Network Data Repository with Interactive Graph Analytics and Visualization", in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[Salt86] Gerard Salton and Michael J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, Inc., New York, NY, USA, 1986.

[Sore48] T. Sørenson, *A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons*, Biologiske Skrifter // Det Kongelige Danske Videnskabernes Selskab, I kommission hos E. Munksgaard, 1948.

[Wang14] P. Wang, B. Xu, Y. Wu and X. Zhou, "Link Prediction in Social Networks: the State-of-the-Art", *ArXiv e-prints*, November 2014.

[Wass94] Stanley Wasserman and Katherine Faust, *Social network analysis: Methods and applications*, vol. 8, Cambridge university press, 1994.

[Yang11] Shuang-Hong Yang, Bo Long, Alex Smola, Narayanan Sadagopan, Zhaohui Zheng and Hongyuan Zha, "Like Like Alike: Joint Friendship and Interest Propagation in Social Networks", in *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pp. 537–546, New York, NY, USA, 2011, ACM.

[Zhan04a] Harry Zhang, "The Optimality of Naive Bayes", in Valerie Barr and Zdravko Markov, editors, *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)*, AAAI Press, 2004.

[Zhan04b] Tong Zhang, "Solving Large Scale Linear Prediction Problems Using Stochastic Gradient Descent Algorithms", in *ICML 2004: PROCEEDINGS OF THE TWENTY-FIRST INTERNATIONAL CONFERENCE ON MACHINE LEARNING. OMNIPRESS*, pp. 919–926, 2004.

[Zhou09] T. Zhou, L. Lü and Y.-C. Zhang, "Predicting missing links via local information", *European Physical Journal B*, vol. 71, pp. 623–630, October 2009.

[Zhu12] Y.-X. Zhu, L. Lü, Q.-M. Zhang and T. Zhou, "Uncovering missing links with cold ends", *Physica A Statistical Mechanics and its Applications*, vol. 391, pp. 5769–5778, November 2012.