FAIR NODE CLASSIFICATION

Πανεπιστήμιο Ιωαννίνων
Τμήμα Μηχανικών Η/Υ & Πληροφορικής
Online Social Networks and Media

Γραπτή Εργασία Εξαμήνου
Γραμμένος Φώτιος

Κεφάλαιο 1: Περίληψη

Με την εξέλιξη της τεχνολογίας τα τελευταία χρόνια η ανθρωπότητα οδεύει σε μία πιο αυτοματοποιημένη καθημερινότητα. Ο άνθρωπος πλέον αναθέτει ένα μεγάλο ποσοστό από τις υποχρεώσεις του σε αυτοματοποιημένα εργαλεία και υπολογιστές. Για να δώσουμε τις ευθύνες σε μία αυτοματοποιημένη υπηρεσία θα πρέπει πρώτα να διασφαλίσουμε κάποια ηθικά ζητήματα. Δεν είναι λίγα τα παραδείγματα όπου αυτοματοποιημένες υπηρεσίες διαχειριζόταν με διαφορετικό τρόπο κάποιες κοινωνικές ομάδες. Για να αποφευχθούν τέτοια λάθη θα πρέπει να αξιολογούμε το πόσο δίκαιος είναι ένας αλγόριθμος.

Κεφάλαιο 2: Εισαγωγή

Για το πόσο δίκαιοι είναι οι αλγόριθμοι, κατηγοριοποιήσεις έχουν αρχίσει να απασχολούν ολοένα και περισσότερο την επιστημονική κοινότητα. Έχουν γίνει πολλές προσπάθειες για τον ορισμό ενός κανόνα ο οποίος θα μπορεί να αξιολογεί τους αλγόριθμους για το πόσο δίκαιοι είναι. Κάτι τέτοιο είναι αρκετά δύσκολο καθώς οι απόψεις διίστανται για τον τρόπο που θα πρέπει να γίνει μία τέτοια προσέγγιση. Σε αυτή την εργασία περιλαμβάνουμε αρκετούς ορισμούς που καθορίζουν τη δικαιοσύνη ενός αλγόριθμου ως προς την κατηγοριοποίηση. Στην περίπτωσή μας εξηγούμε σύμφωνα με ποιες θεωρίες ο αλγόριθμος κατηγοριοποίησης είναι δίκαιος και με ποιες άδικος. Στόχος της συγκεκριμένης εργασίας δεν είναι να θεσπίσει και να καθορίσει ένα τρόπο μέτρησης της δικαιοσύνης, αλλά να εξάγει συμπεράσματα μέσα από μία σειρά πειραμάτων και δοκίμων πάνω σε πραγματικά δεδομένα.

Έχουμε θεσπίσει τους στόχους της συγκεκριμένης εργασίας και τους λόγους για τους οποίους είναι απαραίτητη η μελέτη του συγκεκριμένου θέματος. Στα επόμενα κεφάλαια θα δούμε κάποιες βασικές αρχές και θεωρίες που χρειάζεται να γνωρίζουμε για την κατανόηση της συγκεκριμένης εργασίας. Θα γίνει μία αναφορά για τους λόγους και τον τρόπο που διαλέξαμε τα δεδομένα μας, όπως και το είδος των ορισμών δικαιοσύνης που επιλέξαμε. Έπειτα γίνεται μία αναφορά και περιγραφή της μεθοδολογίας που ακολουθούμε για την υλοποίηση του κώδικα όπου θα μας επιτρέψει και θα μας δώσει όλα τα αποτελέσματα που επιθυμούμε. Στα τελευταία κεφαλαία γίνεται μία αναπαράσταση των πιο σημαντικών αποτελεσμάτων και παρατήρηση για τα πιο ενδιαφέροντα συμπεράσματα που μπορούμε να πάρουμε από την εργασία.

Κεφάλαιο 3: Γνωστικό Υπόβαθρο

Για την κατανόηση της εργασίας χρειάζεται πρώτα να ορίσουμε μερικά βασικά θεμέλια όπου θα χρησιμοποιηθούν στη συνέχεια. Αρχικά οι βάσεις που χρησιμοποιούμε πάρθηκαν έτοιμες από τη μεγάλη γκάμα του stanford. Με κύριο θέμα την δίκαιη κατηγοριοποίηση κόμβων σε κοινωνικά δίκτυα, η επιλογή των βάσεων θα έπρεπε να είναι αυστηρά πάνω σε δίκτυα κοινωνικού περιεχομένου. Οι βάσεις μας παρέχουν τις κατάλληλες προδιαγραφές (protected και unprotected attributes) έτσι ώστε να πραγματοποιήσουμε μία

ομαδοποίηση με κάποια από τα attributes. Ταυτόχρονα θα πρέπει να μας δίνεται η δυνατότητα να παρατηρήσουμε με ποιον τρόπο οι προστατευόμενες πληροφορίες επηρεάζουν την κατηγοριοποιήση που εφαρμόζουμε. Τα προστατευόμενα δεδομένα θα πρέπει να χωρίζονται σε δύο κυρίες κατηγορίες ώστε να παράγουμε ένα διαδικό δίλημμα. Για τους παραπάνω λόγους οι επιλογές βάσεων είναι οι εξής:

Facebook, το συγκεκριμένο dataset παριστάνει ένα γράφο όπου ο κάθε κόμβος είναι ένας χρήστης του Facebook. Κάθε χρήστης περιέχει πληροφορίες για τον ίδιο όπως το φύλο και τον τύπο πτυχίου όπου κατέχει ο κάθε χρήστης. Εδώ ορίσαμε σαν προσωπικά δεδομένα το φύλο του χρήστη και δεδομένα για κατηγοριοποιήση τον τύπο πτυχίου που κατέχει το κάθε προφίλ.

Google Plus, είναι μία βάση δεδομένων, όπως και το Facebook, που περιγράφει κοινωνικά δίκτυα με κάθε κόμβο να είναι το προφίλ ενός χρήστη. Αντίστοιχα το προφίλ περιέχει πληροφορίες για το φύλο του χρήστη, τον τύπο δουλειάς, το μέρος, το μικρό όνομα, το επώνυμο και που πήγε πανεπιστήμιο. Το φύλο του χρήστη ορίζεται ως η προστατευόμενη πληροφορία και εφαρμόζουμε κατηγοριοποιήση ως προς την πληροφορία του χρήστη που αναφέρεται σε ποιο πανεπιστήμιο σπούδασε.

Twitch, είναι μία βάση δεδομένων όπου αντικατοπτρίζει ένα γράφο από διάφορους δημιουργούς περιεχομένου της πλατφόρμας. Κάθε κόμβος συμβολίζει ένα δημιουργό περιεχομένου με τις πληροφορίες που είναι απαραίτητες για το περιεχόμενο που παράγει και τον ίδιο. Για παράδειγμα το μέγεθος των θεατών, εάν το περιεχόμενο του είναι για ενήλικες ή όχι και πολλά άλλα. Για προστατευόμενη πληροφορία ορίζουμε το αν το υλικό που παράγει ο χρήστης είναι για ενήλικες ή όχι. Αντίστοιχα για για την κατηγοριοποιήση θα χρησιμοποιήσουμε την πληροφορία που αναφέρεται στα άτομα που παρακολουθούν τον χρήστη.

Pokec, είναι μία σέρβικη κοινωνική πλατφόρμα σαν το Facebook η οποία συμβολίζεται από έναν γράφο, με κάθε κόμβο να είναι ένα προφίλ της εφαρμογής. Το κάθε προφίλ έχει πολύ σημαντικές πληροφορίες όπως το φύλο του χρήστη, την ηλικία του, τον τόπο καταγωγής του, τα χόμπι του και πολλά άλλα. Σε αυτή τη βάση τα προστατευόμενα δεδομένα του χρήστη είναι το φύλο, ενώ η κατηγοριοποίηση γίνεται ως προς τα είδη των χόμπι του κάθε χρήστη.

Στην περίπτωσή μας εφαρμόζουμε τέσσερις μεθόδους κατηγοριοποιήσης. Πρώτος αλγόριθμος είναι ο Knn, ο όποιος χρησιμοποιεί τις πληροφορίες που παρέχουν οι γειτονικοί κόμβοι (στο πείραμα αυτό ο αριθμός των γειτόνων ισούται με 3). Στην συνέχεια έχουμε τον αλγόριθμο svm, ο οποίος προσπαθεί να βρει την διαχωριστική ευθεία γραμμή που να ξεχωρίζει τις δύο ομάδες δεδομένων. Ο συγκεκριμένος αλγόριθμος υποθέτει ότι οι ομάδες δεδομένων είναι σχετικά ανεξάρτητες. Ωστόσο κάτι τέτοιο δεν ισχυεί για όλες τις βάσεις. Τρίτη μέθοδος κατηγοριοποιήσης είναι ο Decision Tree που έχει εντελώς διαφορετική προσέγγιση με τους προηγούμενους δύο αλγόριθμους, διότι αναπαριστά σε δέντρο τις συνέπειες απο τις αποφάσεις που παίρνει. Έχοντας αντικατοπτρίσει όλα τα πιθανά σενάρια στο δέντρο, επιλέγει την καταλληλότερη προγνωστική τιμή. Τέλος, εφαρμόζουμε

τον αλγόριθμο Bagging. Το διαφορετικό που κάνει αυτός αλγόριθμος κατηγοριοποιήσης είναι ότι σπάει σε υπό κομμάτια τη βάση και εφαρμόζει κατηγοριοποίηση πάνω σε αυτά. Το τελικό αποτέλεσμα βγαίνει από τη μέση τιμή όλων των προηγούμενων αποτελεσμάτων που έχουν προκύψει από τα υπό κομμάτια. Για όλες τις παραπάνω μεθόδους χρησιμοποιούμε τη βιβλιοθήκη sklearn.

Όλοι οι ορισμοί έχουν παρθεί από το paper fairness definitions explained και χρησιμοποιούνται για την αξιολόγηση κάθε αλγορίθμου. Αναφέρονται επιγραμματικά παρακάτω:

Predictive parity. Ένας ταξινομητής ικανοποιεί αυτόν τον ορισμό εάν και οι δύο ομάδες έχουν ίσο PPV. Με άλλα λόγια την ίδια πιθανότητα να αποδοθεί θετική προγνωστική αξία όταν ο υποψήφιος ανήκει πραγματικά στην θετική ομάδα.

False-positive error. Ένας ταξινομητής ικανοποιεί αυτόν τον ορισμό όταν οι δύο ομάδες έχουν ίσο FPR. Δηλαδή την πιθανότητα ενός υποψήφιου πού ανήκει στην αρνητική τάξη να έχει θετική προγνωστική αξία.

False-negative error. Ένας ταξινομητής ικανοποιεί αυτόν τον ορισμό όταν οι δύο ομάδες έχουν ίσο FNR. Στην ουσία την πιθανότητα ενός υποψήφιου πού ανήκει στην θετική τάξη να έχει αρνητική προγνωστική αξία.

Equalized odds. Αυτό ο ορισμός συνδυάζει τους δύο προηγούμενους. Ένας ταξινομητής ικανοποιεί τον ορισμό εάν οι δύο ομάδες έχουν ίσο TPR και ίσο FPR. Μαθηματικά ισοδυναμεί με τον σύνδεσμο της συνθήκης false positive error και false negative error που δίνονται παραπάνω.

Conditional use accuracy equality. Παρόμοια με τον προηγούμενο ορισμό, αυτός ο ορισμός συνδυάζει την ισοδυναμία του PPV και του NPV μεταξύ των δύο ομάδων. Με άλλα λόγια υπολογίζεται η πιθανότητα των υποψηφίων με θετική πρόβλεψη τιμή να ανήκουν πραγματικά στη θετική τάξη και η πιθανότητα των υποψηφίων με αρνητική προγνωστική αξία να ανήκουν πραγματικά στην αρνητική τάξη.

Overall accuracy equality. Ένα ταξινομητή ικανοποιεί αυτός ο ορισμός εάν οι ομάδες έχουν ίση ακρίβεια πρόβλεψης.

Treatment equality. Αυτός ο ορισμός εξετάζει την αναλογία των σφαλμάτων που κάνει ο ταξινομητής παρά στην ακρίβειά του. Ο ταξινομητής ικανοποιεί αυτόν τον ορισμό εάν οι ομάδες έχουν ίση αναλογία ψευδώς αρνητικών και ψευδώς θετικών προβλέψεων.

Κεφάλαιο 4: Προσέγγιση και μεθοδολογία του κώδικα.

Η μεθοδολογία που ακολουθήσαμε στον κώδικά μας περιέχει 4 βασικά βήματα. Το πρώτο είναι η εισαγωγή καθώς και το φιλτράρισμα των γραφών. Φιλτράρισμα ονομάζουμε την μετατροπή τον πολυδιάστατων πληροφοριών του κάθε χρήστη σε δυαδικό σύστημα. Για παράδειγμα στη βάση δεδομένων Pokec ένας χρήστης έχει πολλά χόμπι. Για να γίνει η δυαδική μετατροπή ορίζουμε κάποια χόμπι σαν θετικά άρα αναθέτουμε την τιμή 1 και όλα τα υπόλοιπα σαν αρνητικά , άρα παίρνουν

τιμή 0. Έχοντας φορτώσει τις βάσεις δεδομένων μας στο πρόγραμμά μας, αναθέτουμε τις κατάλληλες τιμές στους κόμβους. Με αυτά τα βήματα προετοιμάζουμε τα δεδομένα για την κατηγοριοποίησή τους. Για την υλοποίηση του επόμενου βήματος χρειάζεται να μεταφέρουμε την πληροφορία στο χώρο των embedding. Κάτι τέτοιο το καταφέρνουμε με τη βοήθεια της βιβλιοθήκης sklearn, όπου κάνουμε χρήση της συνάρτησης node2vec. Το τρίτο βήμα είναι η αρχικοποίηση των αλγόριθμων κατηγοριοποιήσης και ο διαχωρισμός των embedding δεδομένων σε test και training mode. Αφού έχουν ολοκληρωθεί τα παραπάνω βήματα, εκπαιδεύουμε τον αλγόριθμο μας. Στη συνέχεια κατηγοριοπιούμε τα δεδομένα που δεν είχαν λάβει μέρος στην εκπαίδευση του αλγόριθμου. Το επόμενο βήμα είναι η δημιουργία του confusion Matrix (ενός εργαλείου που μας επιτρέπει να κρίνουμε σε ποιο βαθμό η κατηγοριοποίηση είναι σωστή ή όχι). Τέλος, έχοντας όλη την πληροφορία που χρειαζόμαστε μπορούμε πλέον να εφαρμόσουμε τους ορισμούς δικαιοσύνης πάνω στους αλγόριθμους κατηγοριοποίησης.

Κεφάλαιο 5: Αποτελέσματα

Εφαρμόζουμε όλους τους αλγόριθμους κατηγοριοποίησης στις βάσεις μας. Το αποτέλεσμα που παράγεται είναι 4 confusion matrix ανά βαση. Αντί να παρουσιάσουμε και τις τέσσερις βάσεις με παρόμοια αποτελέσματα για τον κάθε αλγόριθμο, θα εστιάσουμε στη Pokec κυρίως για δυο λόγους. Αρχικά είναι η μεγαλύτερη βάση που διαθέτουμε, όποτε μπορεί να αντιπροσωπεύσει μία μεγάλη γκάμα αποτελεσμάτων και πορισμάτων που προκύπτουν από την εφαρμογή των ορισμών δικαιοσύνης. Επίσης, είναι μια βάση με κυρίες κλάσεις σχετικά ισοδύναμες όσο αφορά τον αριθμό συμμετεχόντων σε κάθε μία. Πιο συγκεκριμένα, ο γραφος διαθέτει συνολικά 119,924 εγγραφές.

Στον πρώτο πίνακα βλέπουμε τα αποτελέσματα που παράγει ο αλγόριθμος Knn. Συγκρίνοντας τα PPV για τις δύο ομάδες παρατηρούμε ότι ο αλγόριθμος κατηγοριοποίησης είναι πιο πιθανόν να ορίσει θετική πρόβλεψη τιμή στην ομάδα οπου η προστατευόμενη πληροφορία = 1. Κάτι τέτοιο απορρίπτει τον ορισμό μας οπότε θεωρούμε τον συγκεκριμένο αλγόριθμο μη δίκαιο. Στη συνέχεια παρατηρούμε ότι ο αλγόριθμος έχει μεγαλύτερη πιθανότητα να αναθέσει αρνητική προβλέψιμη τιμή στην ομάδα με προστατευόμενη πληροφορία = 0. Για άλλη μία φορά ο ορισμός μας καταρρίπτεται και θεωρούμε τον αλγόριθμο μας μη δίκιο. Στον τρίτο ορισμό βλέπουμε μία μικρή διαφορά ανάμεσα στις δύο τιμές. Μια τέτοια διαφορά υποδηλώνει την πιθανότητα ο αλγόριθμος να δώσει θετική πρόβλεψη όταν στην πραγματικότητα ο υποψήφιος ανοίγει στην αρνητική ομάδα. Κάτι τέτοιο είναι άδικο. Στον 4ο και 5ο ορισμό μπορούμε αυτομάτως να πούμε ότι ο Knn αποτυγχάνει να είναι δίκαιος διότι σε αυτούς τους ορισμούς υποχρεούνται να ισχύουν δύο ισότητες ανάμεσα στο G0(TPR)=G1(TPR) && G0(FNR)==G1(FNR) και στο G0(PPV)==G1(PPV) && G1(NPV)==G1(NPV) αντίστοιχα. Έχοντας καταρρίψει από τους προηγούμενους ορισμούς το πρώτο μισό της εξίσωσης δεν χρειάζεται να επαληθεύσουμε το

δεύτερο μισό. Με αυτό τον τρόπο μπορούμε να πούμε ότι ο αριθμός αποτυγχάνει και στις δύο περιπτώσεις. Ωστόσο στον ορισμό 6 παρατηρούμε ότι δύο τιμές έχουν μία μικρή διαφορά η οποία μπορεί να θεωρηθεί αμελητέα και για αυτό το λόγο μπορούμε να πούμε ότι ο αλγορίθμος είναι δίκαιος. Τέλος στον ορισμό 7 βλέπουμε μία μεγάλη διαφορά ανάμεσα στα δύο αποτελέσματα και για αυτό το λόγο βγάζουμε το συμπέρασμα ότι ο αλγόριθμος δεν μπορεί να συμβαδίσει με τον ορισμό. Έχοντας μία ολοκληρωμένη εικόνα από τον πίνακα συμπεραίνουμε ότι ο αλγόριθμος Knn τις πιο πολλές φορές είναι άδικος.

Name of Definition	*Protected class =0*	*Protected class =1*	*Result*
Predictive parity	*0.531*	*0.562*	*False*
False-positive error	*0.587*	*0.494*	*False*
False-negative error	*0.417*	*0.485*	*False*
Equalized odds (TPR)	*0.583*	*0.515*	*False*
Conditional use accuracy equality (NPR)	*0.465*	*0.458*	*False*
Overall accuracy	*0.503*	*0.511*	*True*
Treatment equality	*0.811*	*1.208*	*False*

Με την ίδια λογική όπως παραπάνω έτσι και εδώ προκύπτουν τα αποτελέσματα στον πίνακα. Ο συγκεκριμένος πίνακας δείχνει τις τιμές και τα αποτελέσματα για τον αλγόριθμο Baggings. Σε αντίθεση με τον προηγούμενο αλγόριθμο, αυτός επιτυγχάνει να είναι δίκαιος σε δύο ορισμούς, οι οποίοι είναι ο Predictive parity και ο Conditional use accuracy equality (NPR). Μία σημαντική παρατήρηση είναι ότι ο αλγόριθμος Bagging στη βάση Twitch κατέχει όλες τις προϋποθέσεις για τους ορισμούς False-negative error και Equalized odds. Αυτό συμβαίνει μόνο στη συγκεκριμένη βάση και μόνο με το συγκεκριμένο έργο.

Name of Definition	Protected class =0	*Protected class =1*	*Result*
Predictive parity	0.699	*0.703*	*True*
False-positive error	0.583	*0.499*	*False*
False-negative error	0.414	*0.491*	*False*
Equalized odds (TPR)	0.586	*0.509*	*False*
Conditional use accuracy equality (NPR)	0.304	*0.305*	*True*
Overall accuracy	0.535	*0.507*	*False*
Treatment equality	1.642	*2.288*	*False*

Όμοια και ο SVM επαληθεύει τους δύο ίδιους ορισμούς και θεωρείται δικαίως όπως και ο αλγόριθμος Baggings.

Name of Definition	*Protected class =0*	*Protected class =1*	*Result*
Predictive parity	*0.992*	*0.993*	*True*
False-positive error	*0.750*	*0.600*	*False*
False-negative error	*0.416*	*0.495*	*False*
Equalized odds (TPR)	*0.584*	*0.505*	*False*
Conditional use	*0.004*	*0.005*	*True*

accuracy equality (NPR)			
Overall accuracy	*0.582*	*0.505*	*False*
Treatment equality	*89.833*	*147.222*	*False*

Ο τελευταίος αλγόριθμος είναι ο Decision Tree και από τον πίνακα παρακάτω παρατηρούμε όλες τις τιμές που αποδίδονται. Σε αντίθεση με τους προηγούμενους τρεις λογαριασμούς αυτός είναι ο μόνος που βγαίνει δίκαιος σύμφωνα με τον ορισμό Overall Accuracy.

Name of Definition	*Protected class =0*	*Protected class =1*	*Result*
Predictive parity	*0.522*	*0.538*	*True*
False-positive error	*0.581*	*0.500*	*False*
False-negative error	*0.412*	*0.489*	*False*
Equalized odds (TPR)	*0.588*	*0.511*	*False*
Conditional use accuracy equality (NPR)	*0.484*	*0.473*	*False*
Overall accuracy	*0.507*	*0.506*	*True*
Treatment equality	*0.766*	*1.112*	*False*

Τελευταία παρατήρηση για το συγκεκριμένο πείραμα είναι ότι οι ορισμοί Flase-positive error και Treatment equality δεν ικανοποιούνται ποτέ παραμόνο ίσως με την επιλογή κάποιου άλλου αλγόριθμου κατηγοριοποίησης ή με την επιλογή διαφορετικών δεδομένων.

Κεφάλαιο 6: Βιβλιογραφία

- *https://fairware.cs.umass.edu/papers/Verma.pdf*
- *https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html*
- *https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html*
- *https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingClassifier.html*
- *https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html*
- *https://snap.stanford.edu/data/ego-Facebook.html*
- *https://snap.stanford.edu/data/ego-Gplus.html*
- *https://snap.stanford.edu/data/soc-Pokec.html*
- *https://snap.stanford.edu/data/twitch-social-networks.html*