Πολυτεχνείο Κρήτης

Τμήμα Ηλεκτρονικών Μηχανικών & Μηχανικών Η/Υ

Αλγόριθμος Αυτόματης Κατηγοριοποίησης Γραμμών ADSL

Κατατίθεται στο Τμήμα Ηλεκτρονικών Μηχανικών & Μηχανικών Η/Υ για τη μερική ικανοποίηση των προϋποθέσεων λήψης διπλώματος

Εκπόνηση : Κωνσταντίνος Αποστόλου

Εξεταστική Επιτροπή

Καθηγητής Νικόλαος Σιδηρόπουλος (Επιβλέπων)
Καθηγητής Αθανάσιος Λιάβας
Επίκουρος Καθηγητής Γεώργιος Καρυστινός

Οκτώβριος 2010

Πρόλογος

Τα τελευταία χρόνια η διείσδυση της τεχνολογίας DSL και πιο συγκεκριμένα της τεχνολογίας ADSL στην ελληνική αγορά έχει αυξηθεί κατακόρυφα. Ο αριθμός των ADSL συνδέσεων έχει αυξηθεί ραγδαία διεισδύοντας στο 24% των σπιτιών μας. Μάλιστα επειδή οι ADSL συνδέσεις γίνονται κυρίως σε Αστικά Κέντρα έχουμε πολύ μεγαλύτερη πυκνότητα στα κέντρα αυτά, φθάνοντας και ξεπερνώντας ορισμένες φορές και το 50%. Η αυξημένη ζήτηση καθώς και ο μεγάλος όγκος αιτήσεων, δημιουργούν διάφορα προβλήματα στην εύρυθμη λειτουργία των ευρυζωνικών υπηρεσιών. Τα συνηθέστερα προβλήματα είναι η κακή ποιότητα του υπάρχοντος δικτύου τηλεφωνικών καλωδίων, διάφορες πηγές κρουστικού θορύβου (impulse noise), καθώς και το φαινόμενο του crosstalk, το οποίο οφείλεται στην ηλεκτρομαγνητική παρεμβολή μεταξύ γειτονικών ADSL γραμμών. Τα προβλήματα παρεμβολών είναι ισχυρότερα όταν παρέχονται υψηλές ταχύτητες, όπου το ADSL2+ χρησιμοποιεί όλη την ισχύ εκπομπής του για να πετύχει τις υψηλές ταχύτητες.

Όλα τα προαναφερθέντα αποτελούν σοβαρά προβλήματα, όμως από την σκοπιά ενός παρόχου το σημαντικότερο πρόβλημα είναι η ύπαρξη του φαινομένου του crosstalk. Αυτό οφείλεται στο γεγονός ότι κάποιες από τις νέες υπηρεσίες τις οποίες θέλουν να υποστηρίξουν (π.χ. IPTV) απαιτούν απρόσκοπτη επικοινωνία μεταξύ παρόχου και πελάτη. Ποικίλες προσεγγίσεις έχουν ακολουθηθεί στη προσπάθεια να αντιμετωπιστεί το συγκεκριμένο πρόβλημα. Στόχος της παρούσας εργασίας είναι η εξεύρεση γραμμών που πάσχουν από το φαινόμενο crosstalk μέσω μιας αυτοματοποιημένης διαδικασίας.

Εξαιτίας της έλλειψης δεδομένων εκπαίδευσης, χρησιμοποιείται η τεχνική της κατηγοριοποίησης χωρίς επιτήρηση (unsupervised learning / clustering) και πιο συγκεκριμένα μια υλοποίηση του αλγορίθμου k-means. Τα αποτελέσματα του k-means οδηγούνται σε περαιτέρω ανάλυση για ανίχνευση των γραμμών που υπόκεινται στις συνέπειες της ύπαρξης του φαινομένου crosstalk.

Περιεχόμενα

1. Εισαγωγή

2. Παράμετροι ADSL

3. Προαπαιτούμενα - Επισκόπηση

4. Αλγόριθμος Αυτόματης Κατηγοριοποίησης

5. Λεπτομερής Ανάλυση

Κεφάλαιο 1ο - Εισαγωγή

1.1 - Ευχαριστήριο Σημείωμα

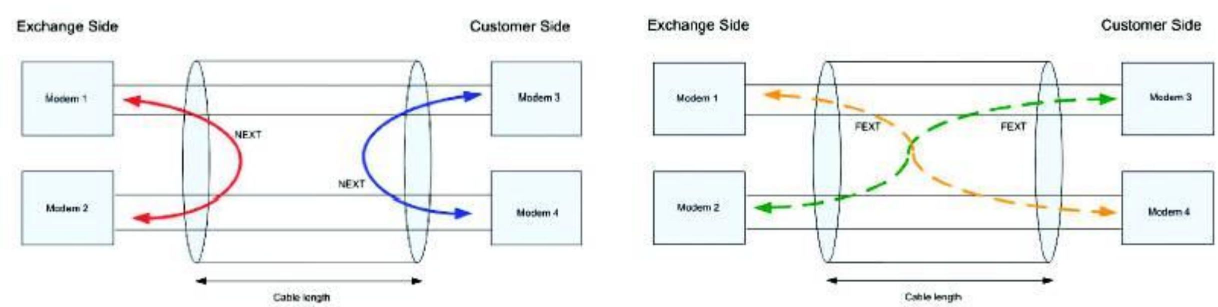
Το θέμα της παρούσας διπλωματικής εργασίας ήταν προϊόν συνεργασίας μεταξύ του επιβλέποντα καθηγητή, κου Σιδηρόπουλου, και του τμήματος ευρυζωνικών συστημάτων της εταιρίας Intracom και πιο συγκεκριμένα του διευθυντή του τμήματος κου Δ. Μήτσαινα και των συνεργατών του κ. Σ. Σπυριδάκη, κ. Μ. Δαλακίδη, κ. Χ. Αθανασίου. Στα πλαίσια αυτής της συνεργασίας, η ομάδα της Intracom μας εφοδίασε με χρήσιμες πληροφορίες, υλικό αλλά και πραγματικά δεδομένα για τη διεκπεραίωση αυτής της εργασίας. Θα ήθελα να τους ευχαριστήσω θερμά για το σύνολο της προσφοράς τους, διότι με την απουσία αυτής δεν θα ήταν δυνατή η περάτωση αυτής της ερευνητικής προσπάθειας.

1.2 - Εισαγωγή στη Τεχνολογία του ADSL

Στα τέλη της δεκαετίας του '90 κατατέθηκαν, από τα 2 μεγαλύτερα Ινστιτούτα Προτύπων στη περιοχή των Τηλεπικοινωνιών, το American National Standards Institute (ANSI) και το International Telecommunications Union (ITU), τα 2 πρότυπα που αποτέλεσαν την απαρχή του ADSL. Το 1998 πρώτο το ANSI κατέθεσε το ANSI T1.413 [1], ενώ ένα χρόνο αργότερα η ITU και πιο συγκεκριμένα το τμήμα ανάπτυξης προτύπων (ITU-T) κατέθεσε το ITU G.992.1 [2]. Τα πρότυπα περιγράφουν τη δυνατότητα χρήσης του υπάρχοντος δικτύου από χάλκινα τηλεφωνικά καλώδια για την ανάπτυξη ενός συστήματος ADSL. Οι αρχικές επιδόσεις δεν ήταν ιδιαίτερα υψηλές αλλά σίγουρα αποτελούσαν ένα τεράστιο άλμα σε σχέση με τις προγενέστερες dial-up υπηρεσίες. Από τότε οι εξελίξεις σε ζητήματα που αφορούν την τεχνολογία ADSL υπήρξαν έντονες, οι διαφοροποιήσεις στις υλοποιήσεις ποικίλες και οι δυνατότητες σημαντικά αυξημένες. Το πλέον σύγχρονο πρότυπο είναι το ITU G.992.5 [3] το οποίο είναι ευρέως γνωστό ως ADSL2+. Οι υψηλές ταχύτητες που προσφέρει και στις 2 κατευθύνσεις του δικτύου (download stream, από το διαδίκτυο στο χρήστη & upload stream, από το χρήστη στο διαδίκτυο) έχουν δώσει δυνατότητα για ανάπτυξη νέων υπηρεσιών που μόλις πριν από λίγα χρόνια φάνταζαν αδύνατες. Όμως οι εξελίξεις και η ιδιαίτερα μεγάλη διείσδυση της τεχνολογίας ADSL στις χώρες που χρησιμοποιείται συνοδεύονται και με τα αντίστοιχα προβλήματα. Η υλοποίηση του ADSL πάσχει από 3 βασικά προβλήματα: a) τα απαρχαιωμένα δίκτυα χάλκινων τηλεφωνικών καλωδίων δημιουργούν συχνά προβλήματα στη σωστή υλοποίηση μιας σύνδεσης αλλά και της εύρυθμης λειτουργίας της, b) η αυξημένη ισχύς εκπομπής του σήματος που απαιτείται στο ADSL2+ επιφέρει ανάλογα αυξημένα προβλήματα με τις παρεμβολές μεταξύ γειτονικών γραμμών, το φαινόμενο που είναι γνωστό ως crosstalk [4] και c) η μεγάλη ευαισθησία της υλοποίησης του ADSL σε πηγές κρουστικού θορύβου (impulse noise) [4].

1.3 - Εισαγωγή Στο Φαινόμενο Crosstalk

Όπως προαναφέρθηκε, τρία είναι τα κύρια προβλήματα της τεχνολογίας ADSL. Από τα τρία αυτά προβλήματα, το δυσχερέστερο και πιο δυσεπίλυτο για τις εταιρίες παροχής διαδικτυακών υπηρεσιών είναι το πρόβλημα του crosstalk. Ως crosstalk ορίζεται η ηλεκτρομαγνητική (Η/Μ) παρεμβολή που δέχεται μια ενεργή γραμμή ADSL από τις "γειτονικές" της ενεργές γραμμές. Τα καλώδια που χρησιμοποιούνται για την μεταφορά της πληροφορίας που σχετίζεται με το διαδίκτυο είναι τα ίδια που χρησιμοποιούνται για την μεταφορά των τηλεφωνικών συνδιαλέξεων. Τα καλώδια αυτά έχουν υποτυπώδη ηλεκτρομαγνητική προστασία, η οποία επιτυγχάνεται με τη τεχνική της συστροφής ζεύγους και επειδή σε κάθε δέσμη (bundle) υπάρχουν εκατοντάδες τέτοια καλώδια είναι φυσικό σήματα υψηλής ισχύος να προκαλούν παρεμβολές στα γειτονικά σήματα. Το crosstalk διακρίνεται σε δύο είδη, το Near End Crosstalk (NEXT) και το Far End Crosstalk (FEXT). Σαν NEXT ορίζεται η παρεμβολή που δέχεται ένας δέκτης Η/Μ σήματος από μια κοντινή πηγή Η/Μ, ενώ ως FEXT ορίζεται η παρεμβολή ενός δέκτη από μια απομακρυσμένη πηγή. Ακολουθούν γραφικά παραδείγματα στο ***γράφημα 1***.



Γράφημα 1. Γραφική απεικόνιση των 2 ειδών crosstalk, NEXT & FEXT
(Γράφημα από: *http://www.exfo.com/en/Library/WaveReview/WRArticle.aspx?Id=143*)

Τα προβλήματα που είναι σε θέση να προκαλέσει το crosstalk κυμαίνονται από απλή αύξηση των εσφαλμένων πακέτων μέχρι και στην απόλυτη αδυναμία της γραμμής να συγχρονιστεί. Η έντονη αύξηση των γειτονικών παρεμβολών μπορεί να οδηγήσει στο "κατακερματισμό" ενός ή περισσοτέρων τόνων, το οποίο μπορεί να οδηγήσει την γραμμή σε αποσυγχρονισμό, δηλαδή σε διακοπή της σύνδεσης του χρήστη με το διαδίκτυο. Κάτι τέτοιο είναι εν γένει απευκταίο, αλλά για τις διάφορες νέες υπηρεσίες που είναι πλέον διαθέσιμες από τους παρόχους, όπως τηλεόραση μέσω διαδικτύου (IPTV) ή τηλεφωνικές κλήσεις μέσω διαδικτύου (Voice Over IP - VoIP) κάτι τέτοιο είναι απαγορευτικό και πρέπει να αποφεύγεται με κάθε κόστος.

Εξαιτίας της σπουδαιότητας του προβλήματος του crosstalk πολλοί ακαδημαϊκοί αλλά και πολλές εταιρείες έχουν προσπαθήσει να προτείνουν κάποιες ικανοποιητικές λύσεις. Οι λύσεις ποικίλουν σε σύλληψη, υλοποίηση αλλά και αποδοτικότητα. Κάποιες προτάσεις, που αφορούν την βραχυπρόθεσμη επίλυση του προβλήματος, αντιμετωπίζουν το φαινόμενο μετά την εμφάνιση του, με τεχνικές όπως το Seamless Rate Adaptation (SRA) ή άλλες αντίστοιχες τεχνικές δυναμικής αναδιανομής του bit-loading στους διαθέσιμους τόνους (dynamic bit re-allocation). Μία τελείως διαφορετική τεχνική, η Δυναμική Διαχείριση Φάσματος (Dynamic Spectrum Management - DSM) [5], μπορεί να προσφέρει πάγιες λύσεις, όμως η υλοποίηση της είναι ιδιαίτερα δύσκολη και δαπανηρή και συνήθως αποφεύγεται. Στα πλαίσια του DSM μπορούν να ακολουθηθούν διαφορετικές τακτικές [6]. Τέλος, υπάρχει και μια άλλη προσέγγιση, η οποία αφορά την "πρόληψη", δηλαδή την διάγνωση του προβλήματος πριν αρχίσει να προκαλεί έντονες διαταραχές στις παραμέτρους μιας γραμμής, χρησιμοποιώντας τεχνικές ανάλυσης τάσεων (trend analysis methods).

1.4 - Στόχος Της Εργασίας

Η παρούσα εργασία ασχολείται με τη παρακολούθηση των ενεργών γραμμών ενός παρόχου και το χαρακτηρισμό τους με βάση διάφορες παραμέτρους. Οι παράμετροι αυτές που χρησιμοποιήθηκαν αφορούν τη λειτουργία της γραμμής και παρακολουθούν διάφορες μεταβλητές που είναι ενδεικτικές για τη ποιότητα της σύνδεσης. Οι παράμετροι ανταλλάσσονται μεταξύ του παρόχου και του χρήστη μέσω του δρομολογητή (router). Αναλυτικότερη παρουσίαση των παραμέτρων καθώς και επεξήγηση του ρόλου της κάθε μιας ακολουθεί στο **εδάφιο 2.2**. Κύριος στόχος της εργασίας ήταν να αναπτυχθεί μια μέθοδος που να μπορεί, αξιοποιώντας ένα μικρό αριθμό από πρωτογενείς παραμέτρους (*primitive parameters*), να είναι σε θέση να χαρακτηρίσει αξιόπιστα αν μια γραμμή λειτουργεί σωστά ή αν αντιμετωπίζει κάποιο πρόβλημα και επειδή έχουμε επικεντρωθεί στο πρόβλημα του crosstalk να εντοπίσει ποιές από τις προβληματικές γραμμές πάσχουν από crosstalk.

Δυστυχώς, η απουσία δεδομένων εκπαίδευσης (training data), δηλαδή στοιχεία γραμμών που είναι γνωστό ότι αντιμετωπίζουν προβλήματα crosstalk, μας ανάγκασε να ακολουθήσουμε μια τεχνική κατηγοριοποίησης χωρίς επιτήρηση (unsupervised learning / clustering) έτσι ώστε να κάνουμε τον διαχωρισμό των γραμμών σε "λειτουργικές" και "προβληματικές". Έτσι τα αποτελέσματα και συμπεράσματα στα οποία έχουμε καταλήξει δεν μπορούν να χαρακτηριστούν ως αδιαμφισβήτητα αλλά οι ενδείξεις είναι ισχυρές για να πείσουν τον αναγνώστη ότι τα αποτελέσματα είναι βάσιμα.

1.5 - Σύνοψη εργασίας

Η εργασία αποτελείται από 5 κεφάλαια. Κάθε ένα από αυτά τα κεφάλαια περιγράφει μια διαφορετική θεματική ενότητα. Ακολούθως παρατίθενται εν συντομία το σύνολο των θεμάτων που πραγματεύεται κάθε κεφάλαιο:

Το παρόν, πρώτο κεφάλαιο αποτελεί μια γενική εισαγωγή στο πρόβλημα που πραγματεύεται αυτή η διπλωματική εργασία.

Στο δεύτερο κεφαλαίο της εργασίας, παρουσιάζονται οι παράμετροι που ήταν διαθέσιμες και χρησιμοποιήθηκαν στη διαδικασία της κατηγοριοποίησης (clustering). Επίσης παρατίθενται συνοπτικά και κάποιες παράμετροι που θα μπορούσαν να φανούν χρήσιμες, αλλά δεν ήταν διαθέσιμες.

Στο τρίτο κεφάλαιο, ορίζονται κάποιες βοηθητικές έννοιες που χρειάζονται για τη κατανόηση των διαδικασιών που ακολουθήθηκαν στην εργασία αλλά και κάποιες ποσότητες που ορίστηκαν για τη διεξαγωγή κάποιων μετρήσεων.

Στο τέταρτο κεφάλαιο, παρουσιάζεται το σύνολο των προσπαθειών που έγιναν με τη χρήση του k-means μέχρι να καταλήξουμε στην μέθοδο που τελικά προτείνεται, με σύντομη αναφορά στις αποτυχημένες προσπάθειες και τους λόγους που αυτές απέτυχαν. Επίσης παρουσιάζεται αναλυτικότερα η τελική μορφή του αλγορίθμου που χρησιμοποιήθηκε για να καταλήξουμε στα τελικά αποτελέσματα.

Στο πέμπτο κεφάλαιο παρουσιάζονται τα αποτελέσματα που εξήχθησαν μέσω της εφαρμογής του αλγορίθμου που αναπτύχθηκε στα πλαίσια της διπλωματικής εργασίας. Επιπλέον, καταγράφονται τα συμπεράσματα στα οποία καταλήξαμε καθώς και προτάσεις για περαιτέρω διερεύνηση γύρω από το θέμα, πιθανές βελτιώσεις για τη περίπτωση που γίνουν διαθέσιμες νέες παράμετροι που μέχρι τη διεξαγωγή της εργασίας δεν ήταν διαθέσιμες καθώς και αξιοποίηση των αποτελεσμάτων για επέκταση σε νέες έρευνες.

Τέλος ακολουθεί ένα παράρτημα στο οποίο δίνεται η βιβλιογραφία και οι πηγές που χρησιμοποιήθηκαν.

Κεφάλαιο 2^{ο} - Παράμετροι ADSL

2.1 - Γενικά

Πριν ξεκινήσει να λειτουργεί μια σύνδεση ADSL, δηλαδή κατά την διάρκεια της διαδικασίας της χειραψίας (**handshake**) [2], διάφορες χρήσιμες πληροφορίες ανταλλάσσονται μεταξύ των modems του παρόχου και του χρήστη. Αξιοποιώντας όλες τις πληροφορίες που ανταλλάσσονται για την ποιότητα του δικτύου, για τις δυνατότητες του καναλιού αλλά και για τυχόν περιορισμούς για μια συγκεκριμένη γραμμή, μεταφερόμαστε σε κατάσταση λειτουργίας (**showtime**) [2]. Επίσης, εκτός από την κατάσταση λειτουργίας, υπάρχει και η δυνατότητα να πραγματοποιηθούν κάποιοι μονομερείς ή διμερείς έλεγχοι (*Single Ended Loop Test - **SELT*** & *Dual Ended Loop Test - **DELT***) οι οποίοι μας δίνουν διάφορες επιπλέον πληροφορίες που δεν είμαστε σε θέση να έχουμε διαθέσιμες όταν η γραμμή είναι ενεργή. Αυτές οι επιπλέον πληροφορίες συνήθως εξυπηρετούν διαγνωστικούς σκοπούς. Η μεταφορά όλης αυτής της πληροφορίας γίνεται μέσω διαφόρων παραμέτρων οι οποίες είναι διαθέσιμες στη πλευρά του παρόχου και μερικές από αυτές είναι διαθέσιμες, μόνο για ανάγνωση (*read-only mode*) στη μεριά του χρήστη. Πολλές από αυτές τις παραμέτρους περιγράφονται μέσα στο πρότυπο του ADSL και είναι υποχρεωτικό να υποστηρίζονται από τους παρόχους, ενώ άλλες έχουν δημιουργηθεί ειδικά από τους παρόχους (*proprietary parameters*) και εξυπηρετούν συγκεκριμένες λειτουργίες, αυξάνοντας τη λειτουργικότητα της γραμμής αλλά και τον βαθμό ελέγχου του παρόχου.

2.2 - Παρουσίαση Παραμέτρων

Το πλήθος των υποστηριζόμενων παραμέτρων, ανάλογα με τον πάροχο και το υλικό (*hardware*) που χρησιμοποιεί, μπορεί και να ξεπεράσει τις εκατό. Οι περισσότερες από αυτές όμως είναι διαθέσιμες μόνο κατά τη διεξαγωγή των διαγνωστικών ελέγχων (**SELT & DELT**) οπότε δεν υπάρχει η δυνατότητα να τις χρησιμοποιήσουμε. Οι παράμετροι που είμαστε σε θέση να χρησιμοποιήσουμε πρέπει να είναι διαθέσιμες σε κατάσταση λειτουργίας. Έτσι ο αριθμός των διαθέσιμων παραμέτρων που μπορούμε να αξιοποιήσουμε περιορίζεται σημαντικά. Στο εδάφιο **2.2.1** παρουσιάζονται οι παράμετροι που ήταν διαθέσιμες σε εμάς στα αρχικά στάδια της εργασίας ενώ στο εδάφιο **2.2.2** παρατίθενται οι επιπλέον παράμετροι που μας έγιναν διαθέσιμες στη φάση της λεπτομερούς ανάλυσης, όπως αυτή ορίζεται στο **κεφάλαιο 5** και για αυτό το λόγο έχουν τιτλοφορηθεί ως "Παράμετροι Λεπτομερούς Ανάλυσης". Η ονομασία των παραμέτρων ακολουθεί ένα πρότυπο. Για όλες τις παραμέτρους οι τέσσερις πρώτοι χαρακτήρες είναι το πρόθεμα "adsl" και έπειτα ακολουθεί ένα περιγραφικό όνομα. Αν η παράμετρος είναι μετρήσιμη και στις δύο ροές πληροφορίας (downstream & upstream) τότε, ανάμεσα στο πρόθεμα "adsl" και στο περιγραφικό όνομα, προστίθενται τέσσερις επιπλέον χαρακτήρες που υποδεικνύουν τη κατεύθυνση της ροής που αφορά η συγκεκριμένη παράμετρος, **A**DSL ***T****ransmission* ***U****nit*

***C**entral Office* - "**ATUC**" όταν πρόκειται για το downstream ή ***A**DSL **T**ransceiver **U**nit **R**emote* - "**ATUR**" για το upstream.

2.2.1 - Βασικές Παράμετροι

Errored Seconds

- ADSLAtucPerfESs : Αθροιστική παράμετρος που συγκεντρώνει τον αριθμό των *downstream* "εσφαλμένων δευτερολέπτων" (errored seconds - ES). Ως ES ορίζεται το δευτερόλεπτο κατά το οποίο συνέβη ένα ή περισσότερα σφάλματα (CRC, LoS, LoF κ.τ.λ.).

- ADSLAturPerfESs : Αθροιστική παράμετρος που συγκεντρώνει τον αριθμό των *upstream* ES.

Uptime

- ADSLLineUpTime : Μετρητής που μετράει το χρόνο που η γραμμή είναι σε κατάσταση *showtime* μετά το τελευταίο συγχρονισμό. Η μέτρηση είναι σε δευτερόλεπτα. Ο μετρητής αυτός μηδενίζεται μετά από έναν αποσυγχρονισμό της γραμμής.

Retrains

- ADSLLineDownCnt : Αθροιστικός μετρητής που απαριθμεί το πλήθος των αποσυγχρονισμών μιας δεδομένης γραμμής. Η παράμετρος αυτή μηδενίζεται περιοδικά.

2.2.2 - Παράμετροι Λεπτομερούς Ανάλυσης

Rate

- ADSLAtucChanCurrTxRate : Στατική παράμετρος που περιέχει το ρυθμό μετάδοσης στον οποίο έχει συγχρονιστεί η γραμμή για την κατεύθυνση από τον ιστό προς το χρήστη (*downstream)*. Η τιμή της υπολογίζεται κατά την διαδικασία της χειραψίας και παραμένει σταθερή μέχρι την αποσύνδεση της γραμμής.

- ADSLAturChanCurrTxRate : Η αντίστοιχη παράμετρος για την άλλη κατεύθυνση *(upstream)***.**

Attainable Rate

- ADSLAtucCurrAttainableRate : Δυναμική παράμετρος η οποία μεταφέρει πληροφορία για τον "*επιτεύξιμο ρυθμό μετάδοσης*" (attainable rate) της κατεύθυνσης από τον ιστό προς το χρήστη. Ως attainable rate ορίζουμε το μέγιστο δυνατό ρυθμό στον οποίο μπορεί να συγχρονίσει η γραμμή μας πληρώντας όλες τις απαιτήσεις της ποιότητας της υπηρεσίας (Quality Of Service - QoS) οι οποίες ορίζονται στο standard του ADSL. Ο υπολογισμός αυτής της παραμέτρου γίνεται καθ' όλη την διάρκεια που η γραμμή είναι ενεργή και δεν γίνεται μόνο μια φορά κατά την χειραψία.

- ADSLAturCurrAttainableRate : Η αντίστοιχη παράμετρος για το upstream.

SNR Margin

- ADSLAtucCurrSnrMgn : Δυναμική παράμετρος που υποδεικνύει το περιθώριο σε dB που έχουμε πάνω από την οριακή τιμή του Signal-To-Noise Ratio (SNR) κατά το οποίο επιτυγχάνουμε σύνδεση με πιθανότητα σφάλματος σε ένα bit (bit error rate - BER) μικρότερη του 10^{-7}. Η παράμετρος αυτή αφορά το downstream.

- ADSLAtucCurrSnrMgn : Η αντίστοιχη παράμετρος για το upstream.

Output Power

- ADSLAtucCurrOutputPwr : Στατική παράμετρος που υπολογίζει τη συνολική εκπεμπόμενη ισχύ στη κατεύθυνση από τον ιστό προς το χρήστη. Ο υπολογισμός της γίνεται κατά τη διαδικασία της χειραψίας και η τιμή της παραμένει σταθερή καθ' όλη τη διάρκεια της σύνδεσης.

- ADSLAtucCurrOutputPwr : Η αντίστοιχη παράμετρος για το upstream.

2.2.3 - Λοιπές Διαθέσιμες Πληροφορίες

Εκτός από αυτές τις παραμέτρους, που σχετίζονται με τη λειτουργία της γραμμής, το script με το οποίο γινόταν η συλλογή των δεδομένων κρατούσε πληροφορίες σχετικά με την ταυτότητα του DSLAM στο οποίο ανήκει η κάθε γραμμή (**DSLAM IP**), τον αριθμό της κάρτας του DLSAM στο οποίο ανήκει η γραμμή (**Card Slot**), τον αριθμό της πόρτας της κάρτας (**Line Index**) έτσι ώστε η κάθε γραμμή να μπορεί να έχει μοναδική ταυτότητα χωρίς να χρειάζεται να καταφύγουμε σε αναζήτηση IP της κάθε γραμμής. Μία τέτοια ενέργεια στη προκείμενη περίπτωση θα ήταν ανούσια καθώς η πλειοψηφία των χρηστών συνδέεται στο διαδίκτυο αξιοποιώντας τη μέθοδο της δυναμικής παραχώρησης IP και έτσι κάθε σύνδεση της ίδιας γραμμής ενδεχομένως συνοδεύεται και από διαφορετική IP. Επιπλέον, αποθηκεύεται και η χρονοσφραγίδα (*timestamp*) της κάθε μέτρησης. Η χρονοσφραγίδα είναι της μορφής "dd/mm/yyyy - hh:mm" δηλαδή η ημερομηνία κατά την οποία έγινε η μέτρηση, ακολουθούμενη από την ώρα κατά την οποία έγινε. (π.χ. 25/06/2010 - 03:44)

2.3 - Μη Διαθέσιμες Παράμετροι

Εκτός από τις παραμέτρους που περιγράφηκαν ανωτέρω, υπήρχαν ποικίλες άλλες παράμετροι που η χρήση τους θα μπορούσε ίσως να είναι πολύτιμη. Δυστυχώς, περιορισμοί στο τρόπο και στο πλήθος των παραμέτρων που μπορούσαμε να έχουμε διαθέσιμες, καθώς και άλλες τεχνικές δυσκολίες που παρουσιάστηκαν στην πορεία της εργασίας δεν μας έδωσαν την ευκαιρία να εξερευνήσουμε τη δυνατότητα να τις εκμεταλλευτούμε για την εξαγωγή κάποιων αποτελεσμάτων ή συμπερασμάτων. Οι παράμετροι αυτές παρουσιάζονται ακολούθως όχι μόνο για λόγους πληρότητας, αλλά για την ανάδειξη της χρησιμότητας τους σε περίπτωση που κάποια μελλοντική προσπάθεια στο συγκεκριμένο πεδίο έρευνας θα μπορούσε να τις έχει διαθέσιμες.

DMT Bin Bits

- ADSLAtucDMTBinBits : Το πλήθος των bits σε κάθε υποφέρον (**subcarrier**). Η μετάδοση των δεδομένων στο καθοδική ροή (downstream) γίνεται σε ένα φάσμα συχνοτήτων (του οποίου τα όρια εξαρτώνται από την έκδοση του ADSL (ADSL/ADSL2/ADSL2+) καθώς και αν η σύνδεση γίνεται με χρήση Public Switched Telephone Network (PSTN) ή Integrated Services Digital Network (ISDN)) το οποίο χωρίζεται σε ζώνες που έχουν πλάτος 4.3125 kHz. Κάθε τέτοιο υποκανάλι, εύρους 4.3125 kHz, μπορεί να μεταφέρει από 1 έως 15 bits ανά σύμβολο ή 0 αν η συγκεκριμένη συχνότητα έχει αποφασιστεί κατά την διαδικασία της χειραψίας να μη χρησιμοποιηθεί.

DMT Bin SNR

- ADSLAtucDMTBinSNR : Παράμετρος στην οποία αποθηκεύεται ένας πίνακας που περιέχει τις μετρήσεις του λόγου ωφέλιμου σήματος προς θόρυβο (SNR) για κάθε subcarrier.

Επιπροσθέτως με τις παραμέτρους που αναφέρονται πιο πάνω, υπάρχουν προδιαγραφές για κάποιες άλλες παραμέτρους, οι οποίες όμως δεν έχουν υλοποιηθεί ακόμα, με τις οποίες θα συλλέγονται πληροφορίες για τη μορφή του φάσματος εκπομπής, την εξασθένιση του σήματος σε κάθε υποφέρον αλλά και μετρήσεις του θορύβου σε κάθε υποφέρον όταν δεν αποστέλλεται κάποιο σήμα.

Κεφάλαιο 3ο -
Προαπαιτούμενα - Επισκόπηση

3.1 - Εισαγωγή

Σε αυτό το κεφάλαιο θα εξηγήσουμε εν συντομία έννοιες και μεθόδους που χρησιμοποιήθηκαν στα πλαίσια της παρούσας διπλωματικής εργασίας. Πιο συγκεκριμένα θα εξηγήσουμε το πρόβλημα της κατηγοριοποίησης (clustering) και έναν συγκεκριμένο αλγόριθμο κατηγοριοποίησης (k-means), τη μέθοδο Monte Carlo καθώς και το σχεδιασμό μίας μετρικής, την οποία χρησιμοποιήσαμε για να μετρήσουμε την ομοιογένεια κάποιων συνόλων.

3.2 - Μέθοδοι Κατηγοριοποίησης

3.2.1 - Εισαγωγή στις Μεθόδους Κατηγοριοποίησης

Ως κατηγοριοποίηση (clustering) ορίζουμε το διαχωρισμό ενός πλήθους αντικειμένων που αντιπροσωπεύονται από ανύσματα σε κατηγορίες (clusters) έτσι ώστε η κάθε κατηγορία να αποτελείται από αντικείμενα που "μοιάζουν" μεταξύ τους. Ο ορισμός και η μέτρηση της ομοιότητας εξαρτάται σε ένα βαθμό από το είδος των δεδομένων. Επίσης η μεθοδολογία που ακολουθείται επηρεάζει τον τρόπο με τον οποίο γίνεται η κατηγοριοποίηση. Υπάρχουν πολλά είδη αλγορίθμων κατηγοριοποίησης. [7], [8] Στα πλαίσια αυτής της εργασίας χρησιμοποιήθηκε ένας διαχωριστικός αλγόριθμος γνωστός ως k-means.

3.2.2 - K-Means

Ο αλγόριθμος k-means είναι ο πλέον διαδεδομένος αλγόριθμος κατηγοριοποίησης.[8] Ο αλγόριθμος λαμβάνει ως είσοδο του τα δεδομένα και το πλήθος των κατηγοριών (clusters) και σκοπός του είναι να τα μοιράσει με τέτοιο τρόπο ώστε να δημιουργηθούν k clusters με τη μέγιστη δυνατή συνάφεια. Ακολούθως παρατίθενται τα βήματα που ακολουθεί ο αλγόριθμος σε μια τυπική υλοποίηση του:

1. Σε πρώτο στάδιο, πρέπει να αρχικοποιηθούν τα δεδομένα στο απαιτούμενο πλήθος clusters. Η συνηθέστερη αρχικοποίηση που ακολουθείται είναι η τυχαία, δηλαδή χωρίς τη χρήση κάποιου συγκεκριμένου κριτηρίου, τα δεδομένα χωρίζονται σε K clusters. Μετά το διαχωρισμό των δεδομένων υπολογίζονται τα κέντρα των clusters (cluster centers) παίρνοντας τον μέσο όρο όλων των δεδομένων που ανήκουν στο ίδιο cluster.

2. Έπειτα, υπολογίζεται η απόσταση κάθε στοιχείου από όλα τα κέντρα και ανατίθεται στο cluster από του οποίου το κέντρο απέχει την μικρότερη απόσταση. Ανάλογα με την υλοποίηση του αλγόριθμου, αλλά κάποιες φορές και το είδος των δεδομένων, διαφορετικές μετρικές για την μέτρηση της απόστασης μεταξύ των δεδομένων μπορούν

να χρησιμοποιηθούν. Χαρακτηριστικά παραδείγματα αποτελούν η Ευκλείδεια απόσταση, η οποία αποτελεί και τη συνηθέστερη επιλογή αν δεν υπάρχει λόγος για την επιλογή κάποιας άλλης μετρικής, η απόσταση συνημίτονου, η οποία υπολογίζεται ως το 1 μείον το συνημίτονο της γωνίας που σχηματίζει το σημείο με το κέντρο του cluster, η απόσταση συσχέτισης, η οποία υπολογίζεται ως το 1 μείον τη συσχέτιση του διανύσματος που δημιουργούν το σύνολο των σημείων που ανήκουν στο ίδιο cluster και τέλος η απόσταση Hamming, η οποία χρησιμοποιείται για δυαδικά δεδομένα και υπολογίζεται ως το ποσοστό των bits που είναι διαφορετικά. Αφού γίνει η ανακατανομή των δεδομένων στα k clusters, επαναϋπολογίζονται τα κέντρα και η διαδικασία επιστρέφει στην αρχή του βήματος 2.

3. Ο αλγόριθμος τερματίζει όταν πλέον κανένα σημείο δεν αλλάξει cluster κατά την εκτέλεση του βήματος 2. Σε αυτή τη περίπτωση ο αλγόριθμος έχει συγκλίνει σε μία λύση.

Ο αλγόριθμος k-means έχει το προτέρημα ότι μπορεί να χειρίζεται στοιχεία οσωνδήποτε διαστάσεων χωρίς να υπάρχει κάποιος περιορισμός και επιπλέον ότι συγκλίνει μονότονα (ως προς το συνολικό τετραγωνικό σφάλμα). Το μειονέκτημα του είναι ότι εν γένει, δεν μπορεί να δοθεί καμία διαβεβαίωση ότι η σύγκλιση του αλγορίθμου θα είναι στη βέλτιστη λύση καθώς υπάρχει και η πιθανότητα η λύση στην οποία κατέληξε να είναι ένα σημείο τοπικού ελάχιστου της ποσότητας που θέλουμε να ελαχιστοποιήσουμε. Η επισήμανση αυτή γίνεται διότι και στα πλαίσια αυτής της εργασίας αντιμετωπίσαμε θέματα με τη σύγκλιση του αλγόριθμου σε διαφορετικές λύσεις, ανάλογα με την αρχικοποίηση.

3.3 - Μέθοδος Monte Carlo

Monte Carlo μέθοδοι (ή Monte Carlo πειράματα) [9] είναι μια κατηγορία υπολογιστικών αλγορίθμων που στηρίζονται σε τυχαίες επαναλήψεις ενός πειράματος για τον υπολογισμό των αποτελεσμάτων τους. Η μέθοδος Monte Carlo φάνηκε χρήσιμη όταν χρειάστηκε να υπολογίσουμε τη πιθανότητα σύγκλισης του αλγορίθμου k-means σε μία συγκεκριμένη λύση. Πιο συγκεκριμένα, χρησιμοποιώντας τα ίδια δεδομένα εισόδου, ο αλγόριθμος **k-means** εκτελείτο *N* φορές, τα αποτελέσματα του αποθηκεύονταν και έπειτα υπολογίζονταν το πλήθος των μοναδικών λύσεων και ο αριθμός εμφανίσεων της κάθε μοναδικής λύσης. Έτσι μπορούσαμε να έχουμε μια καλή προσέγγιση της πιθανότητας σύγκλισης του αλγορίθμου k-means σε μια συγκεκριμένη λύση.

3.4 - Καθορισμός Μετρικής

Στα πλαίσια αυτής της διπλωματικής εργασίας, χρειάστηκε να ορίσουμε μια μετρική που να αντανακλά με ποσοτικό τρόπο την ομοιογένεια και ποιότητα της διαδικασίας της κατηγοριοποίησης. Η μετρική αυτή φάνηκε χρήσιμη, όταν κληθήκαμε να συγκρίνουμε γραφικά όμοια αποτελέσματα που προέκυπταν από την εκτέλεση του k-means με παραπλήσιες εισόδους, έτσι ώστε να καθοριστεί η βέλτιστη λύση. Ακολούθως παρατίθενται τα βήματα που απαιτούνται για τον υπολογισμό της μετρικής:

Βήμα πρώτο

Υπολογίζουμε την συνοχή κάθε cluster, δηλαδή το άθροισμα των τετραγωνικών αποστάσεων των σημείων που καταρτίζουν το cluster από το κέντρο του cluster (Within Cluster Distance - wcd). Το αποτέλεσμα είναι ένα διάνυσμα με τόσες θέσεις, όσα και τα clusters.

$$wcd(j) = \sum_{i\ in\ clust\ j} \|x(i) - c(j)\|^2 ,$$

όπου $\underline{c}(j)$ είναι το κέντρο βάρους του cluster j.

Βήμα δεύτερο

Υπολογίζουμε την ανά-δύο απόσταση των clusters χρησιμοποιώντας ως "αντιπρόσωπο" για κάθε cluster το κέντρο του (Between Cluster Distance - bcd). Το αποτέλεσμα είναι ένας τετράγωνος πίνακας nxn, όπου n είναι ο αριθμός των clusters. Στην θέση bcd(i,j) είναι αποθηκευμένη η απόσταση του cluster i από το j. Ο πίνακας είναι συμμετρικός ως προς την κύρια διαγώνιο του, η οποία προφανώς αποτελείται από μηδενικά στοιχεία.

$$bcd(i,j) = \|c(i) - c(j)\|^2$$

Βήμα τρίτο

Υπολογίζουμε έναν νέο πίνακα, ο οποίος πλέον περιέχει την μετρική υπολογισμένη για κάθε ζευγάρι clusters. Για αυτό το βήμα χρησιμοποιούμε μόνο το άνω τριγωνικό κομμάτι του πίνακα, καθώς αν χρησιμοποιείτο ολόκληρος, η κάθε "συνολική απόσταση" μεταξύ 2 clusters θα υπολογιζόταν 2 φορές. Έτσι το αποτέλεσμα είναι ένας άνω τριγωνικός πίνακας, με τα στοιχεία να δίνονται από τον ακόλουθο τύπο:

$$M(i,j) = \frac{wcd(i) + wcd(j)}{bcd(i,j)}$$

Βήμα τέταρτο

Η συνολική μετρική για το κάθε κατατμημένο σετ εισόδου δίνεται ως το άθροισμα όλων των στοιχείων του πίνακα M. Η λογική πίσω από τον ορισμό της μετρικής μας είναι ότι επιζητούμε να αυξήσουμε την συνοχή του κάθε cluster, δηλαδή να ελαχιστοποιήσουμε την απόσταση των cluster members από το cluster center και ταυτόχρονα να αυξήσουμε όσο το δυνατό περισσότερο την απόσταση μεταξύ των clusters, έτσι ώστε το αποτέλεσμα να είναι ευκρινώς διαχωρισμένο.

Κεφάλαιο 4^{ο} - Αλγόριθμος Αυτόματης Κατηγοριοποίησης

4.1 - Εισαγωγή

Μετά την ανάληψη της ευθύνης διεκπεραίωσης αυτής της ερευνητικής εργασίας από τον γράφοντα, ακολούθησαν οι πρώτες συζητήσεις επί του αντικειμένου μεταξύ των τριών εμπλεκόμενων πλευρών. Οι αρχικές αυτές συζητήσεις μας οδήγησαν στο συμπέρασμα ότι η εργασία θα έπρεπε να διαρθρωθεί σε δύο κύρια μέρη:

a) τον πρωταρχικό διαχωρισμό των γραμμών σε "λειτουργικές" και "προβληματικές" με τη χρήση όσο το δυνατόν λιγότερων παραμέτρων, και αφού είχαμε ικανοποιητικά αποτελέσματα από αυτή τη διαδικασία,
b) προσπάθεια για μια πιο αναλυτική κατηγοριοποίηση των γραμμών, κάνοντας χρήση ενός μεγαλυτέρου αριθμού παραμέτρων, αποσκοπώντας στην ανίχνευση του είδους του προβλήματος που χαρακτηρίζει την κάθε κατηγορία, έτσι ώστε να μπορέσουν οι τεχνικοί της Intracom να προτείνουν λύσεις για την μακροπρόθεσμη επίλυση των προβλημάτων που δημιουργεί το φαινόμενο του crosstalk.

4.2 - Αρχικός Διαχωρισμός

Μετά την οριοθέτηση των στόχων, ακολούθησε η αναζήτηση των κατάλληλων παραμέτρων που θα βοηθούσαν στην εγκυρότερη δυνατή κατηγοριοποίηση των γραμμών. Μετά από διαβουλεύσεις με όλους τους συμμετέχοντες σε αυτή την ερευνητική προσπάθεια, καταλήξαμε ότι οι παράμετροι ***ADSLAtucPerfESs, ADSLAturPerfESs, ADSLLineDownCnt*** & ***ADSLLineUpTime*** θα μπορούσαν να μας προσφέρουν την απαραίτητη πληροφορία ώστε να οδηγηθούμε σε μια αξιόπιστη αρχική κατηγοριοποίηση. Οι τέσσερις αυτές παράμετροι είναι αθροιστικές, γεγονός που δεν μας βοηθάει να τις χρησιμοποιήσουμε άμεσα για εξαγωγή αποτελεσμάτων. Για αυτό το λόγο ακολουθήθηκε μία συγκεκριμένη διαδικασία επεξεργασίας, η οποία εξηγείται στο εδάφιο **4.2.1.**

Όπως έχει προαναφερθεί, δεν υπήρχε η δυνατότητα να έχουμε δεδομένα εκπαίδευσης, δηλαδή δεδομένα που αφορούσαν γραμμές που γνωρίζουμε ότι είναι ή δεν είναι προβληματικές, έτσι ώστε να δούμε τη συμπεριφορά των παραμέτρων κάτω από συνθήκες ύπαρξης ενός συγκεκριμένου προβλήματος, ώστε να είμαστε σε θέση να σχεδιάσουμε ένα σύστημα αναγνώρισης κάποιων αντίστοιχων συμπεριφορών. Έτσι η χρήση τεχνικών εκμάθησης χωρίς επιτήρηση (unsupervised learning) και πιο συγκεκριμένα της τεχνικής της κατηγοριοποίησης (clustering) αποτέλεσε μονόδρομο.

Συνολικά, στα πλαίσια αυτής της διπλωματικής εργασίας, χρησιμοποιήθηκαν τέσσερα σετ δεδομένων. Τα πρώτα δύο, στα οποία βασιστήκαμε για το πρώτο στάδιο της εργασίας, περιείχαν μόνο τις παραμέτρους ***ADSLAtucPerfESs, ADSLAturPerfESs, ADSLLineDownCnt*** & ***ADSLLineUpTime.*** Αυτά τα δύο σετ είχαν δεδομένα για τις ίδιες γραμμές, αλλά σε διαφορετικές χρονικές περιόδους. Η ύπαρξη δύο διαφορετικών σετ βοήθησε στο να αποφευχθεί η προσαρμογή της λύσης σε ένα συγκεκριμένο σετ (over-fitting), αλλά και στην επαλήθευση των αποτελεσμάτων που είχαν προκύψει βάσει του πρώτου σετ και στο δεύτερο. Αργότερα, όταν ασχοληθήκαμε με το δεύτερο στάδιο της εργασίας, είχαμε διαθέσιμα και δύο νέα σετ δεδομένων τα οποία περιείχαν όλες τις παραμέτρους που περιγράφονται στο εδάφιο **2.2.**

4.2.1 - Μετατροπή Βασικών Παραμέτρων

Η παράμετρος ***ADSLLineUpTime*** περιέχει αθροιστικά το χρόνο κατά τον οποίο η γραμμή παρέμεινε σε κατάσταση showtime, μετρούμενο σε δευτερόλεπτα. Ο αθροιστής αυτός, υπό ορισμένες προϋποθέσεις, μηδένιζε και η διαδικασία της άθροισης επανεκκινούσε από την αρχή. Οι προϋποθέσεις αυτές δεν συνδέονται αποκλειστικά με κάποιο αποσυγχρονισμό/ απενεργοποίηση της γραμμής, αλλά και με άλλες συνθήκες που σχετίζονται με το υλικό (hardware) που χρησιμοποιείτο στις υποδομές του παρόχου, από τον οποίο λάβαμε τα δεδομένα μας. Το γεγονός ότι εμφανίζονταν αυτοί οι μηδενισμοί μας παρότρυναν να επεξεργαστούμε και να τροποποιήσουμε τα περιεχόμενα αυτής της αθροιστικής παραμέτρου, έτσι ώστε κάθε συνιστώσα του διανύσματος που προέκυπτε να περιείχε καθαρό χρόνο σε δευτερόλεπτα κατά τον οποίο η γραμμή ήταν ενεργή μεταξύ δύο διαδοχικών μετρήσεων. Στη περίπτωση που η τιμή που συναντούσαμε σε ένα κελί ήταν μεγαλύτερη ίση από την τιμή που περιείχε το προηγούμενο κελί τότε μια απλή αφαίρεση μεταξύ των δύο τιμών μας έδινε το επιθυμητό αποτέλεσμα. Σε αντίθετη περίπτωση, δηλαδή αν το κελί υπό εξέταση είχε μικρότερη τιμή από το προηγούμενο, τότε η τιμή παρέμενε ως είχε, καθώς σε κάποια ενδιάμεση χρονική στιγμή, ο μετρητής είχε μηδενιστεί. Η μεταβλητή η οποία προέκυψε, ουσιαστικά μετρούσε τα δευτερόλεπτα κατά τα οποία η γραμμή ήταν ενεργή μεταξύ δύο διαδοχικών μετρήσεων. Στα πρώτα δύο σετ η διάρκεια μεταξύ δύο μετρήσεων ήταν σχεδόν σταθερή, δηλαδή είχαμε μια σταθερή περίοδο δειγματοληψίας. Η μεταβλητή αυτή ονομάστηκε **dt**.

Για την παράμετρο ***ADSLLineDownCnt*** ακολουθήθηκε ακριβώς η ίδια διαδικασία επεξεργασίας με αυτή που περιγράφηκε για τη παράμετρο ***ADSLLineUpTime*** και το αποτέλεσμα αυτής της επεξεργασίας αποθηκεύθηκε σε μία μεταβλητή που ονομάστηκε **dreset.**

Οι παράμετροι **ADSLAtu*PerfESs** υπέστησαν αντίστοιχη επεξεργασία, μόνο που σε αυτή τη περίπτωση δεν είχαμε μηδενισμό του μετρητή, οπότε δεν χρειάστηκε να μεριμνήσουμε για τη περίπτωση που μια τιμή ήταν μικρότερη της προηγούμενης καθώς τέτοια περίπτωση δεν υπήρχε η δυνατότητα να εμφανιστεί. Έτσι το κάθε στοιχείο του διανύσματος περιείχε τη διαφορά της παρούσας τιμής από την προηγούμενη. Τα αποτελέσματα της επεξεργασίας των δύο παραμέτρων αποθηκεύθηκαν σε μεταβλητές με ονομασίες **atuc_dES** και **atur_dES.**

Κατά σύμβαση, και για τις τέσσερις μεταβλητές, η τιμή του πρώτου στοιχείου τέθηκε ίση με την τιμή που προέκυπτε μετά την εκάστοτε επεξεργασία για το δεύτερο στοιχείο.

4.2.2 - Πρώτες Απόπειρες

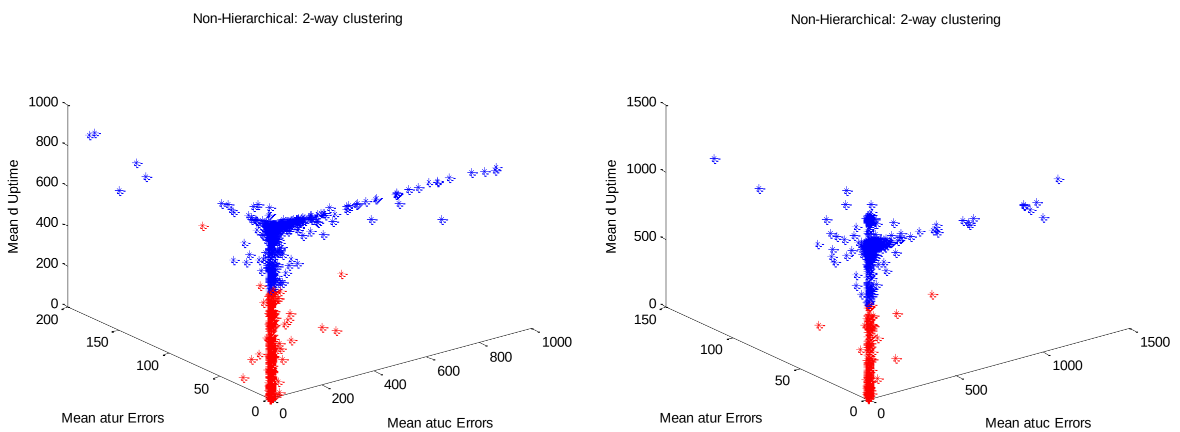
Οι αρχικές απόπειρες επικεντρώθηκαν στη προσπάθεια διαχωρισμού των γραμμών σε 2 κατηγορίες, τις ***υγιείς*** και τις ***προβληματικές.*** Έτσι, χρησιμοποιώντας τις τροποποιημένες παραμέτρους που περιγράφηκαν στο εδάφιο **4.2.1** επιχειρήθηκε, με τη χρήση της έτοιμης υλοποίησης της **Matlab** για τον k-means, ο διαχωρισμός σε 2 σύνολα. Η έτοιμη υλοποίηση της Matlab για τον k-means δέχεται ως ελάχιστο πλήθος ορισμάτων, τα δεδομένα που πρόκειται να κατηγοριοποιηθούν και το πλήθος των clusters στο οποίο αποπειράται να τα διαχωρίσει. Ως είσοδος χρησιμοποιήθηκε ένας πίνακας που αποτελείτο από το μέσο όρο των τεσσάρων μεταβλητών που αναφέρθηκαν ανωτέρω. Έτσι η κάθε γραμμή αναπαρίστατο από τέσσερα στοιχεία, δηλαδή ένα άνυσμα 4x1.

Για να εξετάσουμε τα αποτελέσματα που μας επέστρεψε ο αλγόριθμος θα πρέπει να ορίσουμε κάποιο κριτήριο βάσει του οποίου θα αξιολογήσουμε τη ποιότητα του διαχωρισμού. Πριν αναζητήσουμε ένα ποσοτικό κριτήριο, με τον ορισμό κάποιας μετρικής ομοιογένειας του κάθε αποτελέσματος, προτιμήθηκε η γραφική αναπαράσταση που μας δίνει μια εποπτεία των δεδομένων και της κατανομής τους στο χώρο και βάσει του διαχωρισμού αυτής της αναπαράστασης θα μπορούσαν να εξαχθούν τα πρώτα συμπεράσματα.

Το γεγονός ότι τα δεδομένα μας αναπαριστώνται ως τετραδιάστατα ανύσματα δεν μας επιτρέπει να τα απεικονίσουμε ως έχουν. Είμαστε αναγκασμένοι να ακολουθήσουμε κάποια τεχνική μείωσης των διαστάσεων (dimensionality reduction) από τέσσερις σε - το πολύ - τρεις ώστε να είμαστε σε θέση να τα αναπαραστήσουμε γραφικά και να βγάλουμε τα πρώτα εποπτικά συμπεράσματα.

Επιλέξαμε να προβάλουμε τα δεδομένα μας σε ένα υποχώρο, φυσικά με κάποια απώλεια πληροφορίας, τουλάχιστον ως προς τη γραφική αναπαράσταση. Χρησιμοποιήσαμε τρισδιάστατα γραφήματα σχεδόν στο σύνολο αυτής της εργασίας για την εξαγωγή συμπερασμάτων. Ο άξονας x αντιστοιχεί στο μέσο πλήθος των downstream εσφαλμένων δευτερολέπτων (mean atuc_dES), ο άξονας y στο μέσο πλήθος των upstream εσφαλμένων δευτερολέπτων (mean atur_dES) και ο άξονας z στο μέσο χρόνο ενεργής σύνδεσης της γραμμής ανά χρονοθυρίδα (mean dt). Για τα γραφήματα αυτά επιλέξαμε να αγνοήσουμε τελείως τη πληροφορία που αφορούσε το πλήθος των επανεκκινήσεων (dreset). Η επιλογή αυτή έγινε επειδή η μεταβλητή mean dreset εμφάνισε πολύ μικρότερη διασπορά σε σχέση με τις υπόλοιπες.

Ακολουθεί το αποτέλεσμα της γραφικής αναπαράστασης των διαχωρισμένων δεδομένων για τα δύο διαφορετικά βασικά σετ που ήταν διαθέσιμα.



Γραφήματα 2 & 3. Γραφική αναπαράσταση του διαχωρισμού των δεδομένων σε δύο σύνολα. Το μπλε χρώμα ανταποκρίνεται στις δυνητικά καλές γραμμές ενώ το κόκκινο στις δυνητικά κακές. Αριστερά παρατίθεται το αποτέλεσμα για το πρώτο σετ δεδομένων και δεξιά για το δεύτερο.

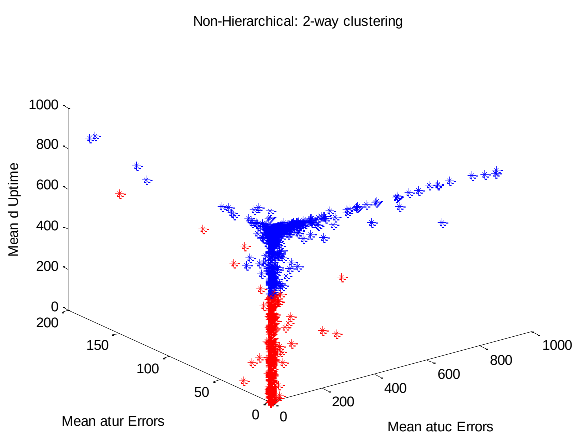
Τα πρώτα συμπεράσματα, βασισμένα στα γραφήματα 2 & 3, δεν ήταν ιδιαιτέρως ενθαρρυντικά, ήταν όμως σημαντικά για τη βελτίωση της μεθοδολογίας που ακολουθήθηκε. Όπως φαίνεται γραφικά, ο διαχωρισμός φαίνεται να επηρεάζεται σε ένα πολύ μεγάλο βαθμό από τις τιμές της παραμέτρου mean dt και φαινομενικά είχαμε ένα διαχωρισμό των δεδομένων στη μέση βάσει της τιμής του μέσου uptime. Αναζητώντας τα αίτια αυτής της συμπεριφοράς του αλγορίθμου, συμπεράναμε ότι ο λόγος ήταν η μεγάλη διαφορά που υπήρχε στο εύρος του συνόλου τιμών των τεσσάρων παραμέτρων. Η παράμετρος m_dt (mean dt) που ήταν και η κυρίαρχη παράμετρος, εκτεινόταν από μηδενικές τιμές μέχρι και τιμές της τάξης του χίλια, ενώ οι 2 μεταβλητές που σχετίζονταν με τα εσφαλμένα δευτερόλεπτα είχαν εύρος από μηδέν μέχρι μερικές εκατοντάδες και τέλος η μεταβλητή m_dreset (mean dreset) είχε εύρος μόλις μία με δύο δεκάδες, με κάτω άκρο το μηδέν.

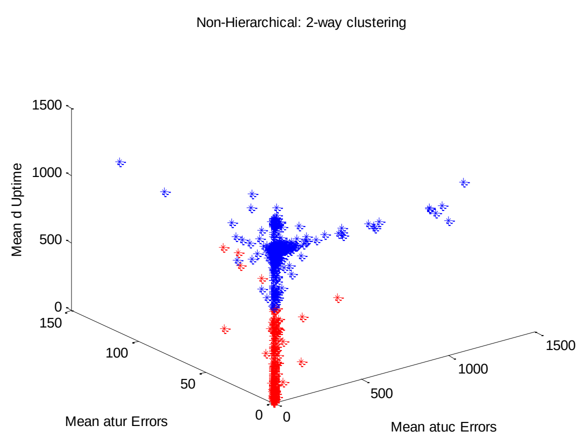
Για να διορθωθεί αυτό το φαινόμενο, καταφύγαμε στην κανονικοποίηση των δεδομένων έτσι ώστε το πεδίο τιμών τους να είναι κοινό, το [0,1]. Αυτό κατέστη δυνατό διαιρώντας την κάθε ακολουθία με τη μέγιστη τιμή της. Παρά αυτή μας την παρέμβαση, η παράμετρος m_dt_norm (η κανονικοποιημένη εκδοχή της m_dt) εξακολουθούσε να είναι κυρίαρχη απέναντι στις υπόλοιπες παραμέτρους. Χαρακτηριστικά αναφέρεται ότι ο αλγόριθμος k-means εξήγαγε τα ίδια ακριβώς αποτελέσματα όταν ως είσοδο δίναμε ένα σετ που αποτελείτο μόνο από την m_dt και όταν το σετ εισόδου αποτελείτο και από τις 4 παραμέτρους.

Εφόσον η κανονικοποίηση απέτυχε να επιφέρει να επιθυμητά αποτελέσματα, και αντιλαμβανόμενοι τα αίτια του φαινομένου, επιχειρήσαμε να τροποποιήσουμε τα βάρη της κάθε μεταβλητής ώστε να μπορέσουμε να μειώσουμε την επιρροή της ***mean dt*** στο αποτέλεσμα αλλά και να ισχυροποιήσουμε τη συμβολή των υπολοίπων. Έτσι η κάθε μεταβλητή απέκτησε δικό της βάρος και μέσω δοκιμών προσπαθήσαμε να βρούμε το καταλληλότερο συνδυασμό που θα μας έδινε ένα αποτέλεσμα διαισθητικά καλύτερο. Το ζητούμενο αυτό αποτέλεσμα ήταν η ενσωμάτωση περισσοτέρων γραμμών με υπερβολικά υψηλές τιμές εσφαλμένων δευτερολέπτων και πολλές επανεκκινήσεις στις δυνητικά κακές γραμμές. Για να πετύχουμε αυτό το αποτέλεσμα, αυξήσαμε το βάρος της μεταβλητής που σχετίζεται με τα ***resets*** κρατώντας παράλληλα σταθερό το βάρος των υπόλοιπων μεταβλητών. Έτσι πλέον το σετ εισόδου του αλγορίθμου k-means μετασχηματίζεται ως ακολούθως:

Από $M = [\overline{\boldsymbol{p_1}}\ \overline{\boldsymbol{p_2}}\ \overline{\boldsymbol{p_3}}\ \overline{\boldsymbol{p_4}}]$ ***(σχέση 1)*** σε $\widehat{M} = [w_1 \overline{\boldsymbol{p_1}}\ w_2 \overline{\boldsymbol{p_2}}\ w_3 \overline{\boldsymbol{p_3}}\ w_4 \overline{\boldsymbol{p_4}}]$ ***(σχέση 2)***

με $\overline{\boldsymbol{p_i}}$ να υποδηλώνεται ένα διάνυσμα στήλης και με w_i να υποδηλώνεται μία βαθμωτή σταθερά. Τα διανύσματα $\overline{\boldsymbol{p_i}}$ ανταποκρίνονται στις μεταβλητές ***mean atuc_dES, mean atur_dES, mean dreset*** και ***mean dt,*** αντιστοίχως. Ένας μεγάλος αριθμός δοκιμών έγιναν, για ποικίλες τιμές των βαρών, όμως τα πιο ελπιδοφόρα αποτελέσματα έφεραν οι δοκιμές κατά τις οποίες θέσαμε σταθερά τα w_1, w_2 και w_4 ίσα με μονάδα και το w_3 να κυμαίνεται μεταξύ των τιμών 2 και 10. Οι δοκιμές αυτές μας οδήγησαν στο συμπέρασμα ότι για μια τιμή του w_3 της τάξης του πέντε, αρκετές γραμμές με υψηλό αριθμό επανεκκινήσεων και εσφαλμένων δευτερολέπτων πλέον τοποθετούνταν στο σύνολο των δυνητικά κακών γραμμών, όχι όμως σε αρκετά ικανοποιητικό βαθμό. Ένα άλλο πρόβλημα που δημιούργησε η χρήση των τροποποιημένων βαρών ήταν στα ποσοστά σύγκλισης του αλγορίθμου k-means στη λύση που θεωρήθηκε βελτιωμένη σε σχέση με τα πρώτα αποτελέσματα στα οποία είχαμε καταλήξει. Όπως αναφέρεται και στο εδάφιο **3.2** ο k-means είναι πιθανό να συγκλίνει σε διαφορετικές λύσεις κατά διαδοχικές εκτελέσεις του, ξεκινώντας από διαφορετικές (π.χ. τυχαίες) αρχικοποιήσεις. Η χρήση των βαρών προκάλεσε την μείωση της σύγκλισης του αλγορίθμου στη καλύτερη λύση από 100% στο 80%-85% επί των συνολικών κλήσεων του. Τα βελτιωμένα αποτελέσματα στα οποία καταλήξαμε, παρατίθενται γραφικά ακολούθως:





Γραφήματα 4 & 5. Γραφική αναπαράσταση του ελαφρά βελτιωμένου αποτελέσματος για τα δύο σετ δεδομένων. Κάποιες γραμμές με υψηλά σφάλματα, υψηλό αριθμό επανεκκινήσεων (δε φαίνεται στα γραφήματα) και όχι τόσο υψηλό uptime έχουν συμπεριληφθεί στις δυνητικά κακές γραμμές.

Όπως βλέπουμε και γραφικά, γραμμές με υψηλό αριθμό σφαλμάτων μεταφέρθηκαν από το "καλό" σύνολο στο "κακό", όμως ακόμα το αποτέλεσμα δεν ήταν πολύ ικανοποιητικό καθώς γραμμές με υπερβολικά υψηλές τιμές σφαλμάτων παρέμεναν στο σύνολο των "καλών" γραμμών. Η απουσία αρκετά καλών αποτελεσμάτων ώθησε την ερευνητική προσπάθεια σε μία σειρά νέων δοκιμών, χρησιμοποιώντας μια διαφορετική τεχνική κατηγοριοποίησης, την ιεραρχική κατηγοριοποίηση.

4.2.3 - Ιεραρχική Κατηγοριοποίηση (Hierarchical Clustering)

Όλες οι αρχικές απόπειρες αλλά και η γραφική αναπαράσταση των γραμμών στο χώρο που επιλέχθηκε, βοήθησαν στη καλύτερη κατανόηση της συμπεριφοράς τους αλλά επίσης οδήγησαν και στην αναζήτηση κάποιων ομάδων / κατηγοριών που ήταν ευδιάκριτες μέσω της γραφικής αναπαράστασης. Παρατηρώντας τις γραμμές, διαπιστώθηκε ότι παρουσιάζονται κάποιες ομάδες γραμμών που μπορούν να ερμηνευθούν και λογικά βάσει των χαρακτηριστικών τους. Οι ομάδες αυτές είναι:

α) η ομάδα του χαμηλού μέσου uptime, δηλαδή οι γραμμές που βρίσκονται συγκεντρωμένες στην ακμή του χώρου, κατά μήκος του z άξονα, με τιμές από μηδέν μέχρι περίπου 400

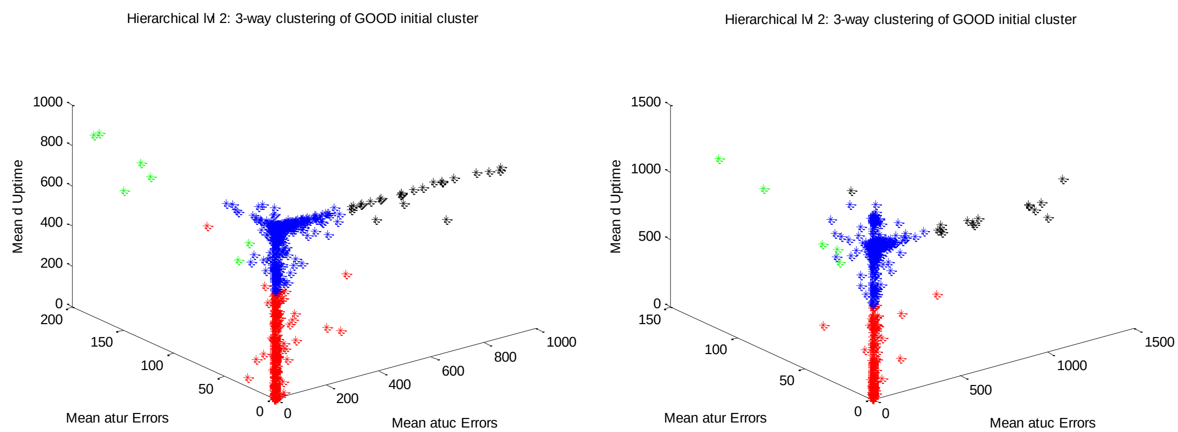
β) η ομάδα υψηλού uptime και χαμηλών σφαλμάτων, που βρίσκεται στο ανώτερο τμήμα του z άξονα με τιμές uptime από 400 και άνω.

γ) η ομάδα υψηλού uptime και υψηλών downstream σφαλμάτων, η οποία βρίσκεται στην x ακμή του χώρου και για τιμές του mean atuc_dES μεγαλύτερες του 200-300.

δ) η ομάδα υψηλού uptime και υψηλών upstream σφαλμάτων, η οποία βρίσκεται στην y ακμή του χώρου και για τιμές του mean atur_dES μεγαλύτερες του 50.

Εξετάζοντας προσεκτικότερα αυτές τις τέσσερις ομάδες μπορούμε να δούμε ότι οι β,γ και δ ομάδες έχουν το κοινό χαρακτηριστικό ότι έχουν υψηλό μέσο uptime και ότι μόνο η ομάδα α έχει χαμηλό uptime. Έτσι ένας αρχικός διαχωρισμός αυτών των δύο "υπερομάδων" θα μπορούσε να πραγματοποιηθεί, κάνοντας χρήση μόνο της μεταβλητής ***mean dt.*** Έπειτα, εξετάζοντας τις γραμμές που χαρακτηρίζονται από υψηλό uptime, μπορούμε να πετύχουμε τον διαχωρισμό τους σε clusters που διακρίνονται από τα χαρακτηριστικά που περιγράφονται για τις ομάδες β,γ και δ κάνοντας χρήση των μεταβλητών που σχετίζονται με τα ES.

Έτσι καταφύγαμε στη χρήση της ιεραρχικής κατηγοριοποίησης για να μπορέσουμε να παράγουμε τις ομάδες που λογικά είχαμε διακρίνει ότι υπάρχουν. Η τεχνική αυτή υπαγορεύει την αρχική κατάτμηση των δεδομένων σε έναν αριθμό ομάδων και έπειτα κάθε μία από αυτές τις ομάδες διαχωρίζονται σε επιμέρους κατηγορίες. Η μεταβλητή ***mean dt*** χρησιμοποιήθηκε για το πρώτο επίπεδο διαχωρισμού σε δύο σύνολα. Το σύνολο που ανταποκρινόταν στις γραμμές με το χαμηλό μέσο uptime αφέθηκε ως είχε, ενώ το δεύτερο σύνολο υπέστη περαιτέρω κατάτμηση. Βάσει του συλλογισμού που περιγράφηκε προηγουμένως, για το δεύτερο επίπεδο επιλέχθηκε το σύνολο να διαχωριστεί σε τρεις νέες υποκατηγορίες. Διεξήχθησαν δοκιμές και με διαφορετικό αριθμό υποσυνόλων αλλά τα καλύτερα αποτελέσματα και πλησιέστερα σε αυτά που στοχεύαμε, προήλθαν από τη κατάτμηση σε 3 υποσύνολα. Ακολουθούν, ενδεικτικά, τα αποτελέσματα για 3-way hierarchical clustering εφαρμοσμένα στα 2 διαθέσιμα data sets. Επειδή, όπως έχουμε αναφέρει και προηγουμένως, ο k-means συγκλίνει σε κάποιο τοπικό ελάχιστο, για τα γραφήματα που ακολουθούν έχουμε κρατήσει την επικρατέστερη λύση για το δεύτερο επίπεδο κατηγοριοποίησης του "αρχικά καλού" συνόλου (γραφήματα 6 & 7). Στο **γράφημα 6** βλέπουμε το διαχωρισμό του "καλού" συνόλου στα 3 υποσύνολα του με τις ***πράσινες*** και τις ***μαύρες*** γραμμές να είναι υποψήφιες προς χαρακτηρισμό ως δυνητικά κακές.

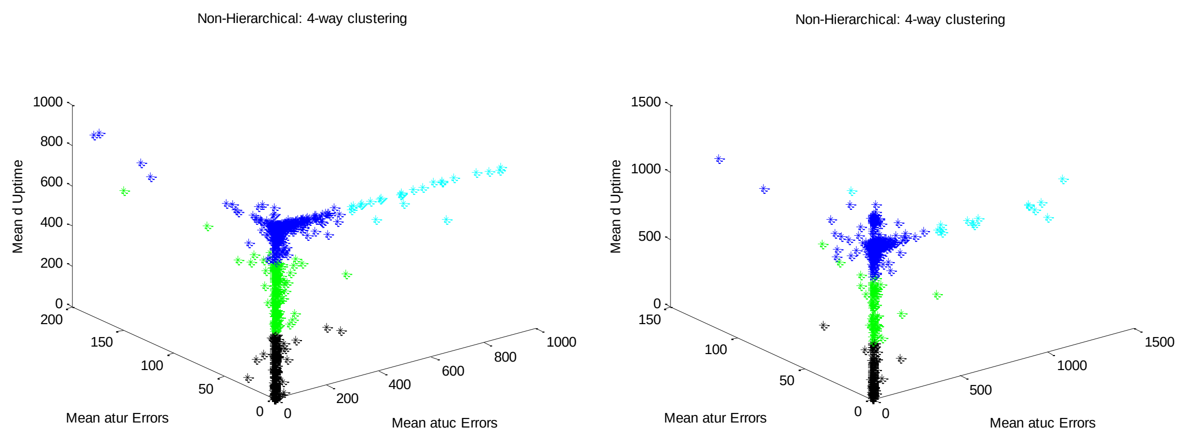


Γραφήματα 6 & 7. Γραφική αναπαράσταση του ιεραρχικού διαχωρισμού για τα δύο διαθέσιμα σύνολα δεδομένων. Παρατηρείται αποτέλεσμα παραπλήσιο με το αναμενόμενο.

Όπως φαίνεται και από τα **γραφήματα 6 & 7** με τη χρήση του ιεραρχικού διαχωρισμού καταφέραμε να δημιουργήσουμε τις ομάδες που λογικά είχαμε εξάγει ότι υπάρχουν. Όμως μετά το πέρας αυτής της προσπάθειας και αφού διαπιστώθηκε ότι υπάρχει η δυνατότητα να διαχωριστούν τα δεδομένα μας σε 4 ευδιάκριτες ομάδες, δοκιμάσαμε να εφαρμόσουμε εκ νέου τον αλγόριθμο k-means στο σύνολο των δεδομένων απαιτώντας τον διαχωρισμό σε 4 clusters, αντί σε 2 όπως είχαμε πράξει στο εδάφιο **4.2.2**.

4.2.4 - Διαχωριστική Κατηγοριοποίηση 4 Συνόλων

Η εμπειρία από τις αρχικές προσπάθειες της κατηγοριοποίησης σε δύο σύνολα, υπαγόρευε ότι θα συναντούσαμε εκ νέου πρόβλημα με την επίδραση της μεταβλητής **mean dt** στο καθορισμό των υποσυνόλων, αλλά για λόγους πληρότητας δοκιμάστηκε ο αλγόριθμος **k-means** με είσοδο το πίνακα M όπως ορίζεται στη ***σχέση 1*** και επιθυμητό πλήθος συνόλων τα τέσσερα. Τα αποτελέσματα, όπως ήταν αναμενόμενο, συνέκλιναν με μεγάλη πιθανότητα σε λύσεις που η επίδραση της **mean dt** είναι κάτι περισσότερο από εμφανής. Ενδεικτικά ακολουθούν οι γραφικές αναπαραστάσεις των εξόδων του αλγορίθμου για τα δύο διαθέσιμα σετ δεδομένων.



Γραφήματα 8 & 9. Γραφική αναπαράσταση της κατηγοριοποίησης τεσσάρων συνόλων για τα δύο διαθέσιμα σύνολα δεδομένων. Τα αποτελέσματα είναι εμφανώς επηρεασμένα από τη μεταβλητή m_dt. Η πιθανότητα σύγκλισης στις λύσεις που παρουσιάζονται ήταν 89.5% για το σετ δεδομένων 1 και 55% για το δεύτερο.

Εξ αιτίας του γεγονότος ότι δεν έχουμε επέμβει εκ νέου στα βάρη της κάθε μεταβλητής που αποτελούν τις στήλες του πίνακα εισόδου του αλγορίθμου, η παράμετρος m_dt (p4) έχει επηρεάσει το αποτέλεσμα εμφανίζοντας 2 ομάδες που έχουν ως κύριο χαρακτηριστικό τη τιμή του uptime και δεν έχει επιτευχθεί ο διαχωρισμός των γραμμών με υψηλά upstream errors σε ξεχωριστή ομάδα.

Για το λόγο αυτό, αναζητήθηκαν τροποποιήσεις του πίνακα εισόδου τέτοιες ώστε να επιτευχθεί ένα ικανοποιητικό αποτέλεσμα, το οποίο να συνοδεύεται και από μία υψηλή πιθανότητα σύγκλισης. Η διαδικασία τροποποίησης των βαρών που περιγράφηκε στο εδάφιο **4.2.2** βασίστηκε πάνω σε διάφορους πειραματισμούς. Σε αυτή την απόπειρα, αφού εντοπίστηκε η μεταβλητή με τη μέγιστη διακύμανση (V_{max}), επιλέξαμε βάρη με τρόπο τέτοιο ώστε η διακύμανση των υπόλοιπων παραμέτρων να ξεκινάει από το ένα δέκατο της V_{max} και να φτάνει, με βήμα $V_{max}/10$, στη τιμή V_{max}. Αναλυτικότερα:

- $v_i = var(p_i) \;\; , \; i = 1{:}4$
- $V_{max} = \max(v_i)$
- $v = k * \frac{V_{max}}{10} \quad , \; k = 1{:}10$
- $w_i = \sqrt{v/v_i} \quad , \; i = 1, 2, 3$
- $w_4 = 1$
- $M_1 = [w_1\,\overline{\boldsymbol{p_1}}\; w_2\,\overline{\boldsymbol{p_2}}\; w_3\,\overline{\boldsymbol{p_3}}\;\; w_4\,\overline{\boldsymbol{p_4}}]$

Τα αποτελέσματα για τις μικρές τιμές των βαρών είναι όμοια με αυτά που παράχθηκαν και από το μη τροποποιημένο σετ, όμως για τιμές του k μεγαλύτερες του τρία, η έξοδος του αλγορίθμου αλλάζει και πλέον τα αποτελέσματα συγκλίνουν σε μία λύση που κρίνεται ιδιαιτέρως ικανοποιητική, τουλάχιστον γραφικά. Ακολούθως παρατίθεται πίνακας με τα αποτελέσματα που προέκυψαν και για τα 2 data sets, για όλες τις τιμές του k, οι πιθανότητες σύγκλισης στην επικρατούσα λύση, καθώς και η γραφική αναπαράσταση των αποτελεσμάτων :

Data Set 1

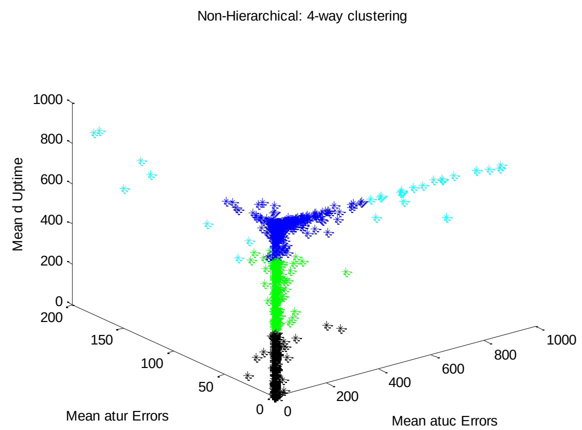
K	1	2	3	4	5	6	7	8	9	10
Convergence	46.20%	53.40%	49.10%	46.10%	76.90%	80.80%	81.40%	84.20%	83.70%	84.60%

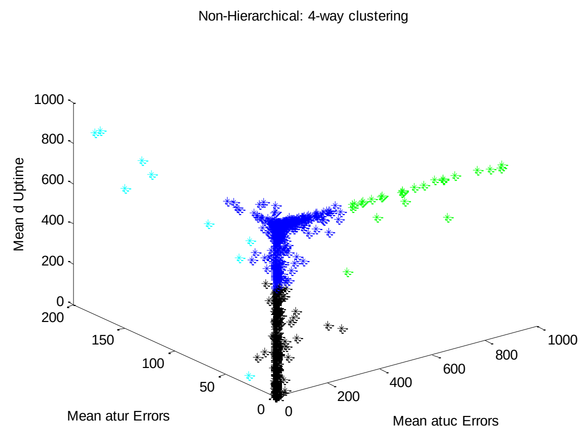
Πίνακας 1 : Παρουσίαση πιθανοτήτων σύγκλισης στην εκάστοτε λύση για διαφορετικές τιμές του k για το σετ 1

Data Set 2

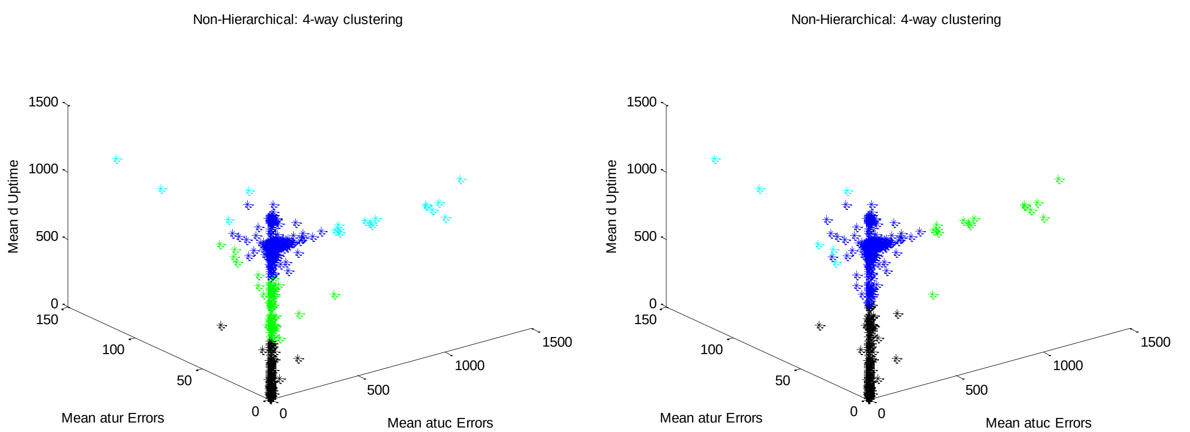
k	1	2	3	4	5	6	7	8	9	10
convergence	83.70%	30.00%	65.60%	70.30%	84.90%	87.10%	91.00%	92.30%	91.10%	91.80%

Πίνακας 2 : Παρουσίαση πιθανοτήτων σύγκλισης στην εκάστοτε λύση για διαφορετικές τιμές του k για το σετ 2





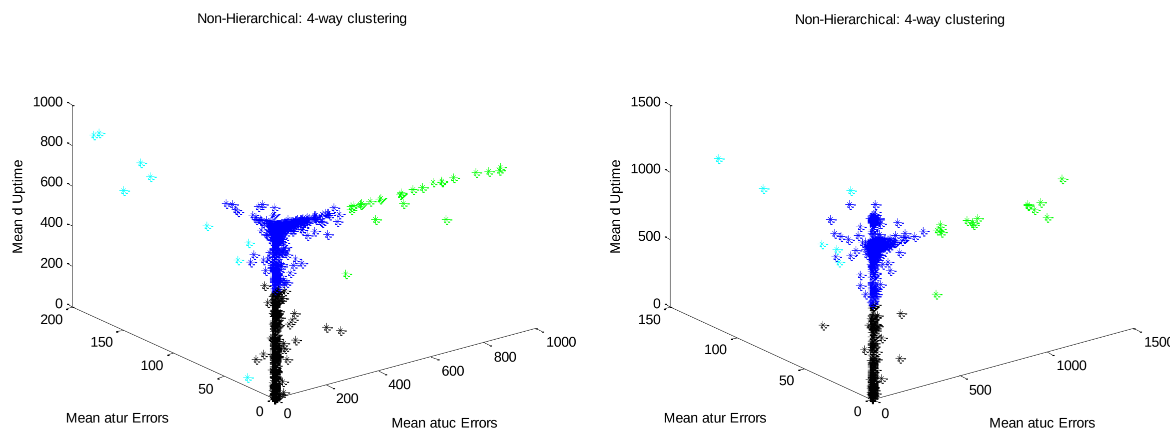
Γραφήματα 10 & 11. Γραφική αναπαράσταση των λύσεων για το σετ δεδομένων 1 και για τιμές του k ≤ 3 (αριστερά) και για τιμές του k ≥ 4 (δεξιά). Το αποτέλεσμα για μικρές τιμές του k δεν κρίνονται επιτυχημένα, όμως μετά τη τιμή 4 η συχνότερα εμφανιζόμενη λύση αλλάζει και κρίνεται ιδιαιτέρως ικανοποιητική.



Γραφήματα 12 & 13. Γραφική αναπαράσταση των λύσεων για το σετ δεδομένων 2 και για τιμές του k ≤ 1 (αριστερά) και για τιμές του k ≥ 2 (δεξιά). Τα αποτελέσματα είναι αντίστοιχα με αυτά που παράχθηκαν για το σετ 1.

Ο βασικός στόχος που είχε τεθεί, δηλαδή η επίτευξη διαχωρισμού των δεδομένων σε τέσσερα υποσύνολα με τα χαρακτηριστικά που περιγράφονται στο εδάφιο **4.2.3**, επιτεύχθηκε. Το μοναδικό μειονέκτημα αυτής της μεθόδου ήταν η σχετικά χαμηλή πιθανότητα σύγκλισης στις κατά τα άλλα ικανοποιητικές λύσεις.

Έτσι κατεβλήθη μια ακόμη προσπάθεια βελτίωσης των ποσοστών σύγκλισης σε αυτά τα ομολογουμένως αρκετά καλά αποτελέσματα. Η μοναδική αλλαγή που μπορούσε να πραγματοποιηθεί σε σχέση με τις συνθήκες διεξαγωγής του προηγούμενου πειράματος ήταν η μέθοδος τροποποίησης του πίνακα εισόδου. Ποικίλες προσπάθειες και συνδυασμοί οδήγησαν στην ακόλουθη μορφή, $M_2 = [w\,\overline{\boldsymbol{p_1}}\quad w\,\overline{\boldsymbol{p_2}}\quad w\,\overline{\boldsymbol{p_3}}\quad \overline{\boldsymbol{p_4}}]$, δηλαδή στο καθορισμό κοινού βάρους για τις τρεις πρώτες μεταβλητές και βάρος μονάδα για τη μεταβλητή p_4 (mean dt). Ένα εύρος από τιμές δοκιμάστηκαν για το βάρος ***w***, μέχρι που ο δευτερεύων στόχος που είχε τεθεί, επετεύχθη και αυτός με τη σειρά του. Για τιμή ***w = 4*** καταφέραμε να έχουμε ίδια αποτελέσματα με αυτά που εξάγαμε προηγουμένως μόνο που επιπλέον είχαμε πετύχει και βελτιωμένη πιθανότητα σύγκλισης σε αυτές τις λύσεις. Έτσι για το σετ δεδομένων 1 η πιθανότητα σύγκλισης στη λύση που ακολουθεί γραφικά ήταν 98.9% ενώ το αντίστοιχο ποσοστό για το δεύτερο σετ ήταν 99.8%. Ακολούθως παρουσιάζονται τα γραφικά αποτελέσματα για τη συγκεκριμένη τιμή του *w:*



Γραφήματα 14 & 15. Γραφική αναπαράσταση των λύσεων για τα δύο σετ δεδομένων και για τιμές του w = 4. Αριστερά παρατίθεται η λύση για το σετ 1 και δεξιά η αντίστοιχη για το σετ 2. Όπως ευδιάκριτα διακρίνεται, οι λύσεις είναι ίδιες με τις αντίστοιχες του πρώτου ορισμού του σετ εισόδου, όμως η πιθανότητα σύγκλισης σε αυτές τις λύσεις είναι αισθητά βελτιωμένες.

Μετά από όλους τους πειραματισμούς και τις δοκιμές που περιγράφονται ανωτέρω, επετεύχθη ο στόχος που είχε τεθεί, οι γραμμές χωρίστηκαν σε τέσσερα σύνολα που έχουν κοινά χαρακτηριστικά. Όμως η αξιολόγηση που είχε γίνει μέχρι στιγμής ήταν καθαρά εποπτική, με τη βοήθεια των γραφικών αναπαραστάσεων. Μέχρι αυτό το σημείο της εργασίας δεν είχε χρησιμοποιηθεί κάποια μαθηματική μέθοδος μέτρησης της υπεροχής μιας λύσης έναντι κάποιας άλλης. Για το λόγο αυτό, σε συνεργασία με τον επιβλέποντα καθηγητή, ορίστηκε μια μετρική που ποσοτικοποιεί την ποιότητα κάθε λύσης. Η μετρική που έχει ορισθεί στο εδάφιο **3.4** βοήθησε στο να εξετάσουμε αν ο χαρακτηρισμός των αποτελεσμάτων ως "ικανοποιητικά" με βάση τις γραφικές αναπαραστάσεις δύναται να υποστηριχθεί και από μια αριθμητική εκτίμηση. Όπως αναφέρεται και στο εδάφιο 3.4 η μετρική έχει ορισθεί με τρόπο τέτοιο ώστε να ελαχιστοποιείται όταν μια λύση αποτελείται από ομογενή σύνολα, δηλαδή τα δεδομένα που αποτελούν κάθε cluster βρίσκονται σε μικρή απόσταση μεταξύ τους, αλλά και ταυτόχρονα τα σύνολα μεταξύ τους απέχουν όσο το δυνατόν περισσότερο.

Βάσει αυτής της μετρικής, έγιναν μετρήσεις για τα αποτελέσματα στα οποίο καταλήξαμε στο εδάφιο **4.2.4** και για όλες τις τιμές παραμέτρων που εξετάσαμε (τη παράμετρο *k* στο πρώτο ορισμό και τη *w* στο δεύτερο). Ακολούθως δίνονται πίνακες με συγκεντρωμένα τα αποτελέσματα αυτών των μετρήσεων.

k	1	2	3	4	5	6	7	8	9	10
data set 1	257.40	292.30	338.11	109.87	116.13	122.95	130.11	137.48	145.00	152.62
data set 2	206.40	65.02	66.50	69.68	73.57	77.83	82.31	86.91	91.61	96.38

Πίνακας 3: Τα αριθμητικά αποτελέσματα της μετρικής για το data set 1 & 2 όταν ήταν ορισμένα ως $[\boldsymbol{w_1\,\overline{p_1}\;\; w_2\,\overline{p_2}\;\; w_3\,\overline{p_3}\;\; w_4\,\overline{p_4}}]$ για όλες τις τιμές του k

w	1	2	3	4	5	6	7	8	9	10
data set 1	202.07	538.51	125.39	163.42	213.52	275.70	349.11	433.45	531.18	638.42
data set 2	123.87	61.18	74.92	97.83	128.32	165.08	209.42	260.60	316.53	380.48

Πίνακας 4: Τα αντίστοιχα αποτελέσματα της μετρικής για το data set 1 & 2 όταν ήταν ορισμένα ως $[\boldsymbol{w\,\overline{p_1}\;\; w\,\overline{p_2}\;\; w\,\overline{p_3}\;\; \overline{p_4}}]$ για όλες τις τιμές του w

Τα ποσοτικά αποτελέσματα δεν συμφωνούν πλήρως με τα συμπεράσματα που καταλήξαμε χρησιμοποιώντας τις γραφικές απεικονίσεις των λύσεων που μας παρείχε ο αλγόριθμος, αλλά και η πιθανότητα σύγκλισης σε αυτές τις λύσεις, όμως δεν είναι και εντελώς αντιφατικά. Όπως φαίνεται, για το πρώτο σετ δεδομένων, και συγκρίνοντας τους δύο διαφορετικούς τρόπους που ορίσαμε το σετ εισόδου, το σετ M_1 και για συντελεστή k = 4 φαίνεται να παράγει καλύτερη λύση απ' ότι το αντίστοιχο M_2 και για τιμή w = 3, όμως η διαφορά μεταξύ των δύο τιμών της μετρικής μπορεί να θεωρηθεί αμελητέα έναντι της συντριπτικής υπεροχής της πιθανότητας σύγκλισης του δεύτερου έναντι του πρώτου, με 46.10% *(πίνακας 1)* για τον πρώτο και 98.9% για τον δεύτερο. Όταν εξετάζεται το δεύτερο σετ δεδομένων, τότε τα αποτελέσματα στα οποία καταλήξαμε εποπτικά είναι ακόμα περισσότερο συμβατά με αυτά που προκύπτουν με τη χρήση της μετρικής, δηλαδή ο δεύτερος τρόπος ορισμού υπερέχει του αντίστοιχου πρώτου και από ποσοτικής άποψης, έχοντας μικρότερη μετρική. Αυτό που προκαλεί εντύπωση, είναι ότι και για τους δυο ορισμούς, το ελάχιστο εμφανίζεται για τιμές των συντελεστών ίσες με 2. Γραφικά τα αποτελέσματα για τιμές από 2 και άνω είδαμε (γράφημα 12-13) ότι ταυτίζονται και το μόνο που τροποποιείται είναι η πιθανότητα σύγκλισης στη λύση. Έτσι μπορούμε να πούμε ότι για συγκεκριμένες τιμές των συντελεστών, δημιουργείται ένα trade off μεταξύ της τιμής της μετρικής και της πιθανότητας σύγκλισης. Για τιμές άνω του 2 παρατηρούμε αύξηση της μετρικής, όμως ταυτόχρονα η πιθανότητα σύγκλισης αυξάνεται. Έτσι η τελική επιλογή του συντελεστή πρέπει να γίνει εξετάζοντας ταυτόχρονα και τις δύο αυτές παραμέτρους για να επιτευχθεί ο καλύτερος δυνατός συνδυασμός. Για το λόγο αυτό, χρησιμοποιώντας τον δεύτερο τρόπο ορισμού (M_2), επιλέχθηκε η τιμή του συντελεστή w=4 για την οποία έχουμε ελαφρά μεγαλύτερη τιμή της μετρικής, 97.83 έναντι 61.18 (w=2), αλλά επιτυγχάνουμε σημαντικά καλύτερη πιθανότητα σύγκλισης, 99.8% έναντι 72.4%. Τα αποτελέσματα για το πρώτο τρόπο ορισμού είναι σαφώς υποδεέστερα καθώς όταν η μετρική παίρνει την ελάχιστη τιμή της, 65.02 για k = 2, η πιθανότητα σύγκλισης είναι 40% *(πίνακας 2)*, ενώ η μέγιστη πιθανότητα σύγκλισης που ήμασταν σε θέση να πετύχουμε ήταν 92,3% για k = 8 με τη μετρική για τη συγκεκριμένη τιμή του k να είναι 86.91.

Κεφάλαιο 5ο - Λεπτομερής Ανάλυση

5.1 - Εισαγωγή

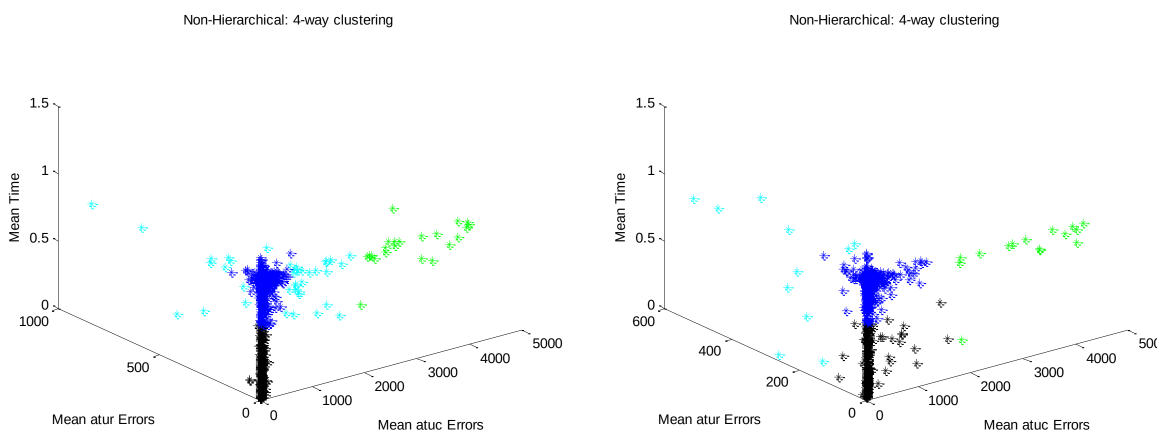
Μετά την επίτευξη του διαχωρισμού των γραμμών σε κατηγορίες που φαίνεται να χαρακτηρίζονται από κοινά στοιχεία, το επόμενο στάδιο της εργασίας προέβλεπε την αναλυτικότερη και λεπτομερέστερη ανάλυση αυτών των συνόλων και τη διερεύνηση αν κάποιο από αυτά εμπεριείχε τις γραμμές που πάσχουν από το φαινόμενο του crosstalk. Για να πραγματοποιηθεί αυτή η λεπτομερής ανάλυση, απαιτούνται σετ δεδομένων με περισσότερες παραμέτρους ανά γραμμή, έτσι ώστε να μπορέσουμε να διαγνώσουμε αν διαφαίνονται ενδείξεις ύπαρξης του φαινομένου.

Αρχικά είχε αποφασιστεί, σε συνεννόηση με την ομάδα υποστήριξης της Intracom, να τους αποσταλεί μια λίστα με όλες τις υποψήφιες "κακές" γραμμές, και από τα τρία σύνολα που μας είχαν προκύψει κατά τη διαχωριστική κατηγοριοποίηση τεσσάρων συνόλων, και για τις εν λόγω γραμμές να συλλέγαμε επιπλέον παραμέτρους για να υπήρχε η δυνατότητα μιας νέας ανάλυσης, σε μεγαλύτερο βάθος, ώστε να εντοπίζονταν οι γραμμές που έπασχαν από το φαινόμενο του crosstalk. Όμως κάτι τέτοιο τελικά δεν κατέστη δυνατό, και έτσι τα αναλυτικά σετ δεδομένων, τα οποία μας αποστάληκαν για περαιτέρω διερεύνηση, συμπεριελάμβαναν το σύνολο των γραμμών τις οποίες είχαμε εξετάσει και στο πρώτο στάδιο της εργασίας. Έτσι επιπλέον των παραμέτρων που είχαμε διαθέσιμες αρχικά, οι οποίες περιγράφονται στο εδάφιο **2.2.1**, στο νέο σετ δεδομένων συμπεριλαμβάνονταν και οι παράμετροι που περιγράφονται στο εδάφιο **2.2.2.**

5.2 - Τροποποίηση Αλγορίθμου

Εξαιτίας του γεγονότος ότι τα νέα σετ δεδομένων περιείχαν το σύνολο των γραμμών και όχι μόνο τις δυνητικά κακές του πρώτου σκέλους της εργασίας, κρίθηκε απαραίτητο και λογικό να εφαρμόσουμε τον αλγόριθμο που χρησιμοποιήθηκε για να εξαχθούν τα αποτελέσματα της πρώτης φάσης της εργασίας και στα νέα σετ δεδομένων. Επειδή στα αναλυτικά σετ συλλέγαμε περισσότερες παραμέτρους ανά γραμμή, η περίοδος δειγματοληψίας, όχι μόνο είχε αυξηθεί αλλά σταμάτησε να είναι και (σχεδόν) σταθερή, συνθήκη η οποία ίσχυε για τη περίοδο δειγματοληψίας των δύο βασικών σετ δεδομένων. Για το λόγο αυτό, δεν κατέστη δυνατό να εφαρμοστεί αυτούσιος ο αλγόριθμος που είχε χρησιμοποιηθεί πρωτύτερα, διότι εμφανίστηκαν επιπλοκές εξαιτίας της μεταβλητής **m_dt,** που σχετίζεται, εμμέσως, με τη περίοδο δειγματοληψίας.

Η τροποποίηση που έγινε ήταν ιδιαιτέρως απλή και δεν επηρέασε την έκβαση του πειράματος. Στη θέση της μεταβλητής **m_dt** που χρησιμοποιείτο μέχρι εκείνη τη στιγμή στο σετ εισόδου του αλγορίθμου ***k-means*** αλλά και για τη γραφική απεικόνιση του αποτελέσματος, δημιουργήθηκε μια άλλη μεταβλητή, η οποία ονομάστηκε **time.** Η μεταβλητή αυτή εμπεριείχε έναν καθαρό αριθμό, ο οποίος δινόταν από το λόγο του ενεργού χρόνου μιας γραμμής προς το πραγματικό χρόνο που μεσολαβούσε μεταξύ δύο διαδοχικών μετρήσεων. Για την δημιουργία αυτής της μεταβλητής χρησιμοποιήθηκαν η παράμετρος ***ADSLLineUpTime*** αλλά και το ***timestamp.*** Η χρονική διαφορά μεταξύ των δύο χρονοσφραγίδων μας έδινε τη χρονική περίοδο που μεσολάβησε μεταξύ των δειγμάτων, και αφού μετατρεπόταν σε δευτερόλεπτα, η διαίρεση της με τη μεταβλητή ***dt,*** η οποία ήταν επίσης εκφρασμένη σε δευτερόλεπτα, μας έδινε ένα ποσοστό που αντιστοιχούσε στο ποσοστό του χρόνου κατά τον οποίο η γραμμή παρέμενε ενεργή. Μέσω αυτής της τροποποίησης επετεύχθη η ανεξαρτητοποίηση από τη περίοδο δειγματοληψίας. Τα **γραφήματα 16** & **17** αναπαριστούν τα αποτελέσματα της εφαρμογής του τροποποιημένου αλγορίθμου στα δύο νέα σετ δεδομένων.



Γραφήματα 16 & 17. Γραφική αναπαράσταση της εφαρμογής του τροποποιημένου αλγορίθμου στα δύο νέα αναλυτικά σετ δεδομένων. Όπως φαίνεται τη θέση της m_dt στον άξονα z παίρνει η νέα μεταβλητή mean Time της οποίας οι τιμές κυμαίνονται από 0 έως 1.

Εξετάζοντας τα παραπάνω γραφήματα μπορεί να υποστηριχθεί ότι τα αποτελέσματα του αλγορίθμου ήταν ανάμεικτα. Η γενική συμπεριφορά κρίνεται ικανοποιητική και αναμενόμενη, όμως υπάρχει μια μικρή παραφωνία στα μέχρι τώρα ιδιαιτέρως ικανοποιητικά αποτελέσματα. Στο **γράφημα 16,** στο οποίο παρουσιάζονται τα αποτελέσματα για το πρώτο αναλυτικό σετ δεδομένων παρατηρείται μια ομάδα γραμμών που αναμενόταν να ανήκε στο σύνολο των γραμμών που έχουν σκιαγραφηθεί με το γαλάζιο χρώμα και όχι σε αυτές που ανήκαν στο πράσινο σύνολο. Επίσης η πιθανότητα σύγκλισης σε αυτή τη λύση ήταν σαφώς μικρότερη σε σχέση με αυτά τα οποία είχαν παρουσιαστεί στο εδάφιο **4.2.4** καθώς το ποσοστό ήταν της τάξης του 91.5%. Όμως αυτή η παραφωνία μπορεί να αποδοθεί στο ίδιο το σετ δεδομένων και κάποιες ασυνήθιστες τιμές που ενδεχομένως να περιείχε καθώς τα αποτελέσματα για το δεύτερο αναλυτικό σετ είναι ακριβώς αυτό που αναμενόταν και μάλιστα το ποσοστό σύγκλισης κυμαινόταν μεταξύ 99.9% και 100%. Το ενθαρρυντικό στοιχείο, όπως διαφαίνεται και στο εδάφιο που ακολουθεί, είναι ότι η εν γένει συμπεριφορά και τα στοιχεία που χαρακτηρίζουν τα 4 υποσύνολα που δημιουργούνται και για τα δύο σετ δεδομένων φαίνεται να είναι παρόμοια και να παρουσιάζουν μια συνέπεια.

5.3 - Παρουσίαση Χαρακτηριστικών Των Ανιχνευμένων Κατηγοριών

Το παρόν εδάφιο είναι αφιερωμένο στη παρουσίαση των παραμέτρων που χαρακτηρίζουν τις γραμμές που ανήκουν στο ίδιο σύνολο. Η ανάλυση αυτή αποσκοπεί στην ανεύρεση του συνόλου εκείνου που είναι περισσότερο πιθανό να εμπεριέχει τις γραμμές που πάσχουν από το φαινόμενο του crosstalk. Πριν όμως παρουσιαστούν τα αποτελέσματα πρέπει να εξεταστεί το είδος των διαταραχών που θα μπορούσε να επιφέρει το φαινόμενο του crosstalk και πως αυτές φανερώνονται μέσα από τις παραμέτρους που έχουμε διαθέσιμες.

5.3.1 - Επενέργειες Του Φαινομένου Crosstalk

Στη τεχνολογία ADSL χρησιμοποιείται η διαμόρφωση διαχωρισμού συχνοτήτων (Frequency Division Multiplexing - FDM) ώστε να δημιουργηθούν Ν ανεξάρτητα υποκανάλια τα οποία χρησιμοποιούνται για την εκπομπή του σήματος. Το κάθε υποκανάλι χαρακτηρίζεται από μια φέρουσα συχνότητα. Η ύπαρξη του φαινομένου του crosstalk μπορεί να πλήξει είτε μερικές από τις συχνότητες είτε και το σύνολο τους. Πιο ευάλωτες είναι οι υψηλές συχνότητες καθώς το σήμα σε αυτές εξασθενεί περισσότερο. Η ηλεκτρομαγνητική παρεμβολή αυξάνει τον θόρυβο στις συχνότητες τις οποίες επηρεάζει και έτσι ο σηματοθορυβικός λόγος (SNR) μειώνεται, γεγονός που δημιουργεί προβλήματα στην εύρυθμη λειτουργία της γραμμής τα οποία μπορούν να έχουν ποικίλες συνέπειες.

Χωρίς τη δυνατότητα να έχουμε πληροφορίες για τη φασματική πυκνότητα ισχύος αλλά ούτε και για το πλήθος των bit που χρησιμοποιούνται ανά φέρουσα, το οποίο είναι συνάρτηση του σηματοθορυβικού λόγου σε κάθε υποκανάλι, έπρεπε να βρεθεί ένας τρόπος έμμεσης ανίχνευσης των συνεπειών του φαινομένου έτσι ώστε να μπορούμε να είμαστε σε θέση να το αναγνωρίσουμε.

Εξετάζοντας τα δεδομένα τα οποία είχαμε στην διάθεση μας, υποθέσαμε ότι γραμμές οι οποίες πάσχουν από crosstalk, χαρακτηρίζονται από χαμηλό μέσο όρο *επιτεύξιμου ρυθμού μετάδοσης,* παράμετρος που είναι δυναμική οπότε μας δίνει αρκετά αξιόπιστες πληροφορίες, αλλά και χαμηλές τιμές του περιθωρίου σηματοθορυβικού λόγου. Επιπλέον ενδεικτική μπορεί να χαρακτηριστεί και η μεταβλητή ***uptime*** αλλά και η, προερχόμενη από αυτή, ***time.***

5.3.2 - Παρουσίαση Αποτελεσμάτων

Για κάθε μεταβλητή παρατίθενται πέντε ιστογράμματα τα οποία αντιστοιχούν στα τέσσερα clusters τα οποία έχουν προκύψει από τη διαχωριστική κατηγοριοποίηση τεσσάρων συνόλων καθώς και ένα συγκεντρωτικό που αφορά το σύνολο των γραμμών. Στα ιστογράμματα που σχετίζονται με τα clusters της κατηγοριοποίησης έχουν δοθεί αριθμητικές ταυτότητες τις οποίες θα χρησιμοποιήσουμε για το υπόλοιπο αυτής της εργασίας. Στο νούμερο μηδέν - 0 έχει αντιστοιχηθεί το "μπλε" σύνολο που συναντάται στο **γράφημα 16** για το πρώτο αναλυτικό σύνολο δεδομένων και το αντίστοιχο "μπλε" σύνολο στο **γράφημα 17** για το δεύτερο αναλυτικό σετ. Κατά πλήρη αντιστοιχία έχουν δοθεί οι αριθμητικές ταυτότητες 1, 2 και 3 για τα "μαύρο", "πράσινο" και "γαλάζιο" σύνολα που συναντώνται στα **γραφήματα 16 & 17**.

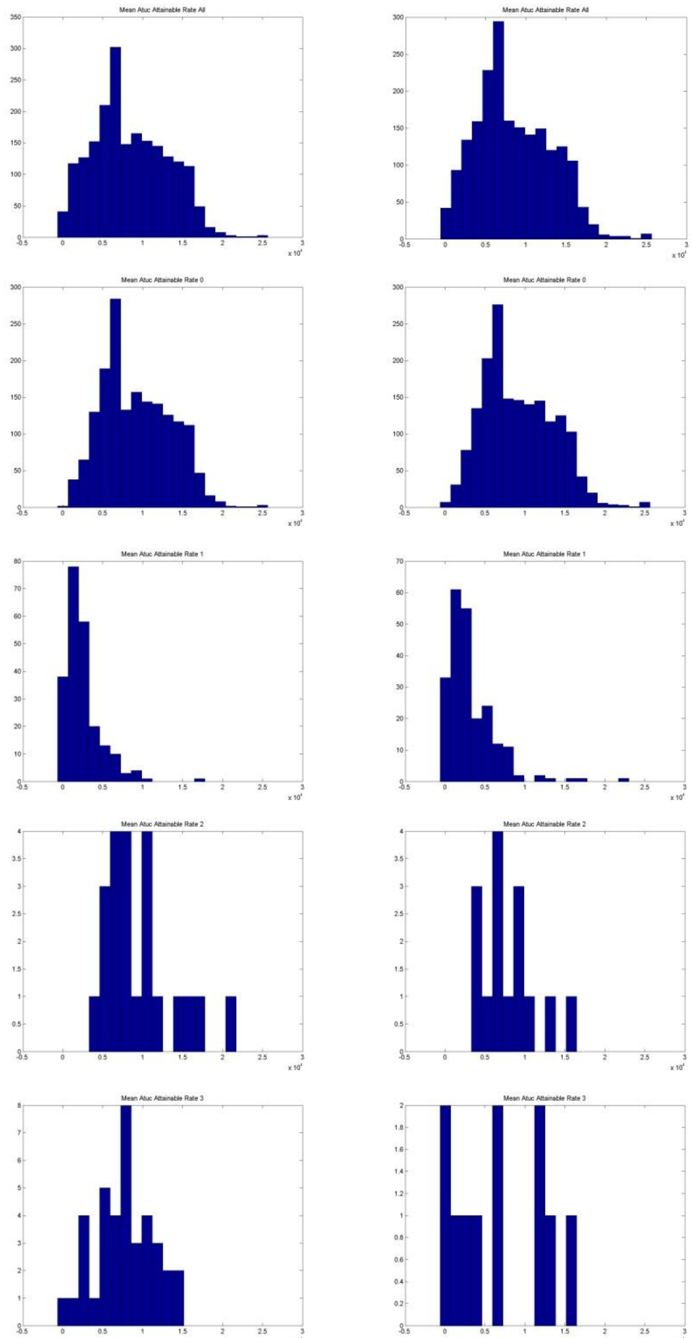
Τα ιστογράμματα που ακολουθούν αφορούν τις εξής μεταβλητές:

- **Mean atuc attainable rate :** μέσος downstream επιτεύξιμος ρυθμός
- **Mean atur attainable rate :** μέσος upstream επιτεύξιμος ρυθμός
- **Mean atuc SNR margin :** μέσο downstream περιθώριο SNR
- **Mean atur SNR margin :** μέσο upstream περιθώριο SNR
- **Mean time :** μέσος λόγος ενεργού προς συνολικού χρόνου ανά χρονοθυρίδα

Ακολούθως παρατίθενται ιστογράμματα που αφορούν όλες τις παραμέτρους που περιγράφονται παραπάνω, για τα δύο αναλυτικά σετ δεδομένων, πίνακες με συγκεντρωμένες τις μέσες τιμές των τεσσάρων clusters για την κάθε μεταβλητή καθώς και η ανάλυση αυτών των δεδομένων.

- **Downstream Attainable Rate**

Data Set 1 **Data Set 2**



Η συγκεκριμένη μεταβλητή μας δίνει τη δυνατότητα να εξετάσουμε τον επιτεύξιμο ρυθμό μετάδοσης για κάθε ένα από τα σύνολα στα οποία έχουμε καταλήξει.

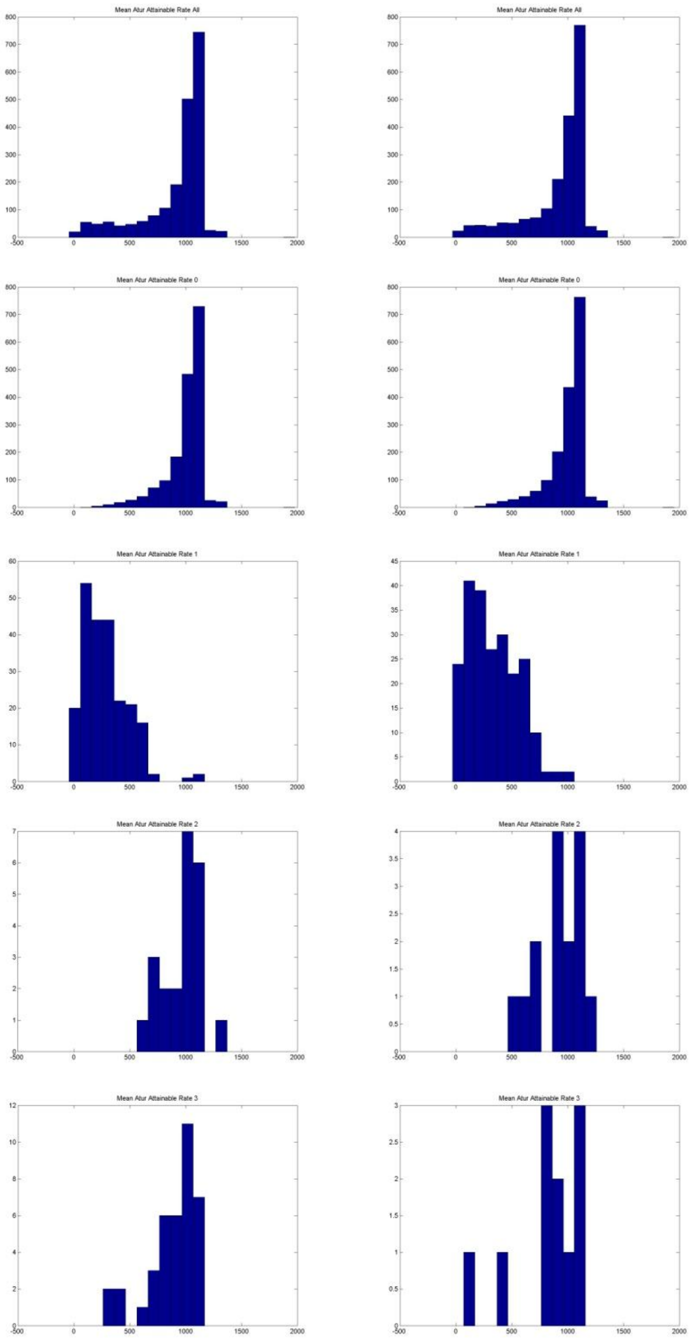
Στη πρώτη γραμμή έχουμε το συγκεντρωτικό ιστόγραμμα, δηλαδή αυτό που περιέχει όλες τις γραμμές, το οποίο μας βοηθάει να αποκτήσουμε μια γενική ιδέα για τη κατανομή των μέσων επιτεύξιμων ρυθμών.

Ακολουθούν τα ιστογράμματα που αφορούν τα τέσσερα σύνολα στα οποία έχουν αποδοθεί οι αριθμητικές ταυτότητες από 0 έως 3. Σε αυτά, δίνεται η δυνατότητα στον αναγνώστη να παρατηρήσει τη κατανομή των μέσων επιτεύξιμων ρυθμών ανά σύνολο.

Όπως φαίνεται και γραφικά, αλλά και όπως καταδεικνύουν και οι **πίνακες 5 & 6**, στο *σύνολο 1* είναι συγκεντρωμένες η πλειονότητα των γραμμών που έχουν χαμηλό επιτεύξιμο ρυθμό και μόνο ελάχιστες απομακρύνονται από το μέσο όρο του συνόλου. Όπως αναφέρεται και προηγουμένως, η χαμηλή δυνατότητα σε επιτεύξιμο ρυθμό είναι ένα στοιχείο που υποδηλώνει κακή ποιότητα καναλιού και κατά πάσα πιθανότητα ύπαρξη παρεμβολών που μπορεί να είναι μάρτυρας ύπαρξης crosstalk.

- **Upstream Attainable Rate**

Data Set 1 **Data Set 2**



Αντίστοιχες παρατηρήσεις μπορούν να γίνουν και για αυτή τη μεταβλητή, η οποία επίσης δίνει τον μέγιστο επιτεύξιμο ρυθμό, αλλά για τη κατεύθυνση από το χρήστη προς το διαδίκτυο.

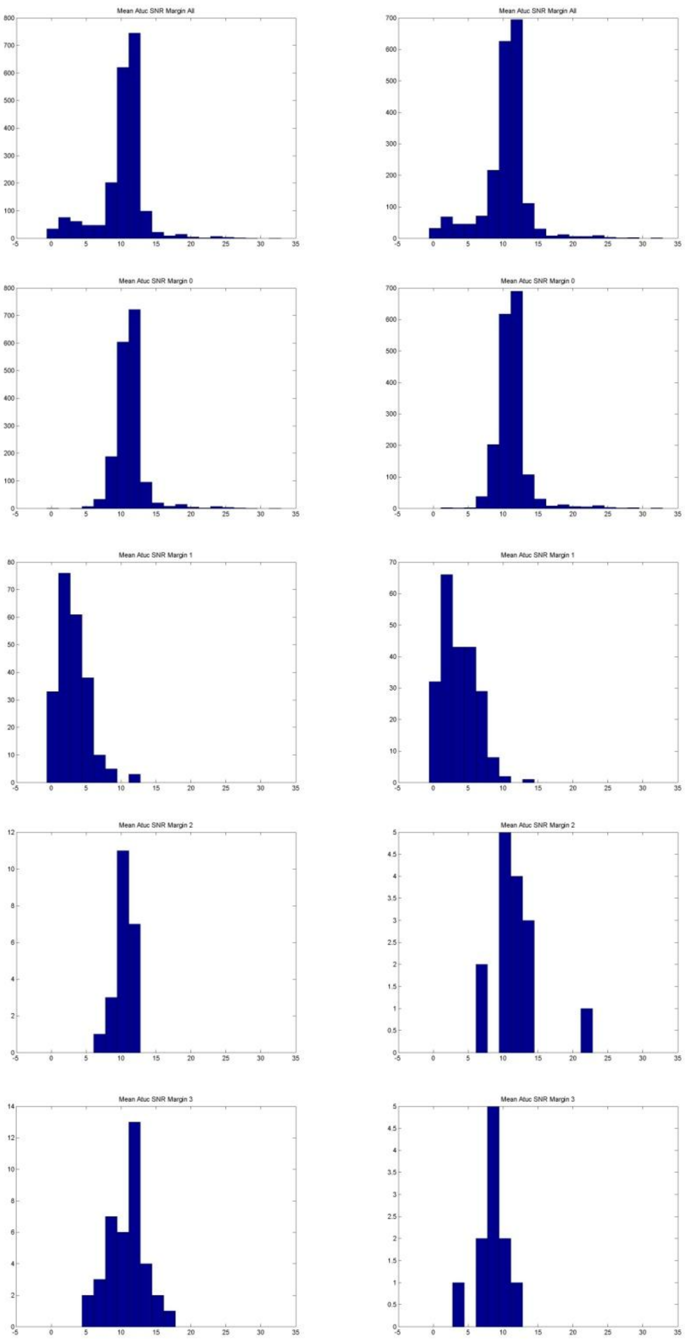
Όπως και για τη μεταβλητή ***downstream attainable rate,*** η πλειονότητα των γραμμών με χαμηλό επιτεύξιμο ρυθμό είναι συγκεντρωμένες στο *σύνολο 1* και για τα δύο αναλυτικά σετ δεδομένων.

Έτσι μπορεί να ειπωθεί ότι, όσον αφορά το διαχωρισμό γραμμών που έχουν περιορισμένες δυνατότητες στον επιτεύξιμο ρυθμό, εξαιτίας δυσμενών συνθηκών στο κανάλι, έχει γίνει ένα καθοριστικό και αρκετά πετυχημένο βήμα.

Με την βοήθεια των υπόλοιπων ιστογραμμάτων γίνεται δυνατή η ενίσχυση του ισχυρισμού ότι η πλειονότητα αυτών των γραμμών που έχουν διαχωριστεί, πάσχουν από ύπαρξη ηλεκτρομαγνητικών παρεμβολών και ενδεχομένως του φαινομένου crosstalk.

- **Downstream SNR Margin**

Data Set 1 **Data Set 2**



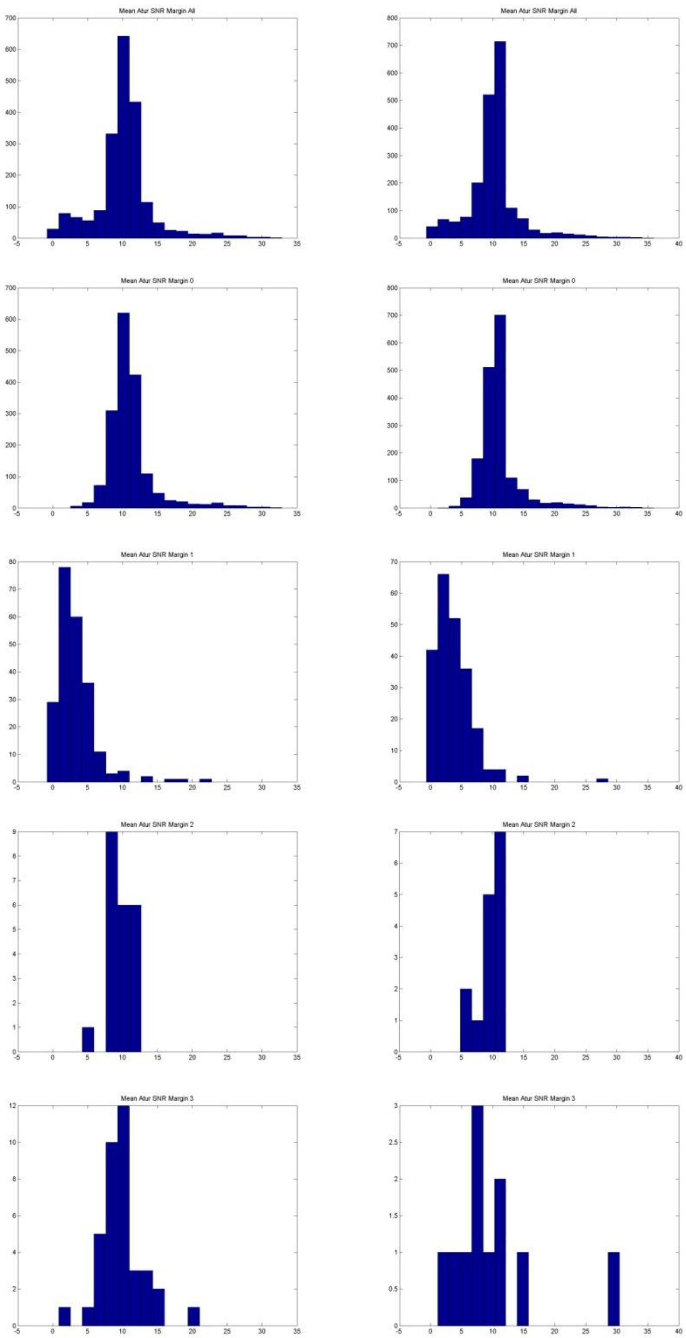
Η μεταβλητή ***SNR margin*** εμπεριέχει το περιθώριο σε dB που έχει η γραμμή που εξετάζεται σε σχέση με το ελάχιστο απαιτούμενο SNR ώστε να επιτυγχάνεται σύνδεση με πιθανότητα σφάλματος μικρότερη ίση του 10^{-7}.

Για το λόγο αυτό, γραμμές με χαμηλό ***SNR Margin*** καταδεικνύουν αυξημένη ύπαρξη θορύβου, καθώς για σταθερή ισχύ εκπομπής, μια αύξηση στα επίπεδα του θορύβου, μειώνει το σηματοθορυβικό λόγο ο οποίος πλησιάζει περισσότερο στην οριακή τιμή του **SNR** κατά την οποία έχουμε επίτευξη του επιθυμητού **BER.**

Όπως ξεκάθαρα φαίνεται από τα ιστογράμματα που παρατίθενται παραπλεύρως, οι γραμμές που ανήκουν στο *σύνολο 1* χαρακτηρίζονται από ιδιαιτέρως χαμηλές τιμές σηματο-θορυβικού περιθωρίου, γεγονός που ενισχύει τους ισχυρισμούς μας για ύπαρξη παρεμβολών, και κατ' επέκταση φαινομένου **crosstalk** στις γραμμές που το απαρτίζουν. Σε πλήρη αντίθεση όλα τα υπόλοιπα σύνολα χαρακτηρίζονται από αρκετά καλές τιμές του **SNR margin,** καθώς τιμές ανώτερες των 10 dB (για το downstream) εγγυώνται κατ' ελάχιστον μια σταθερή σύνδεση.

- **Upstream SNR Margin**

Data Set 1 **Data Set 2**

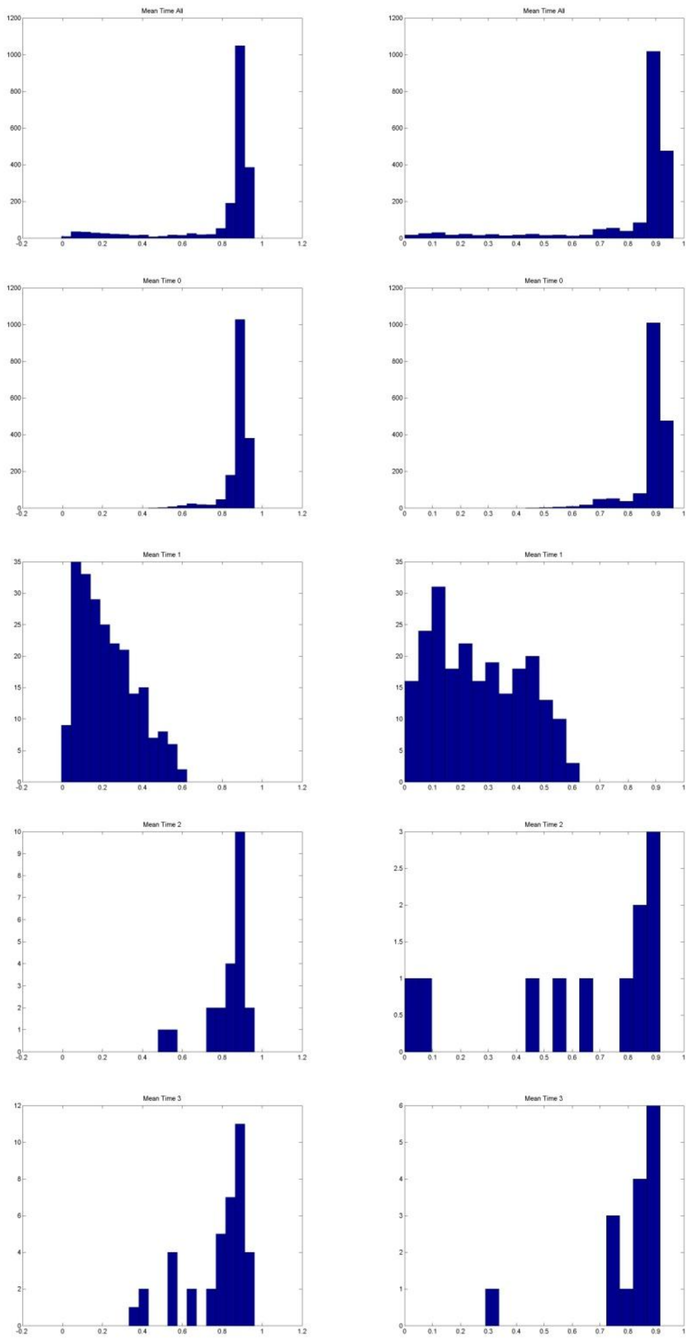


Εξετάζοντας και αυτή τη μεταβλητή, παρατηρούνται αντίστοιχα δεδομένα με τις προηγούμενες μεταβλητές. Το μεγαλύτερο ποσοστό των γραμμών που χαρακτηρίζονται από χαμηλό σηματοθορυβικό περιθώριο είναι συγκεντρωμένες στο *σύνολο 1,* ενώ τα υπόλοιπα σύνολα παρουσιάζουν εν γένει καλύτερες συνθήκες.

Όπως και για το **Downstream SNR Margin**, έτσι και για το **Upstream SNR Margin**, οι χαμηλές τιμές είναι ένδειξη χαμηλής ποιότητας καναλιού και υψηλών επιπέδων θορύβου.

- **Time**

Data Set 1 **Data Set 2**



Η μεταβλητή **time,** όπως έχει αναφερθεί και προηγουμένως, είναι μια τεχνητή μεταβλητή που μας δίνει το λόγο του χρόνου κατά τον οποίο η γραμμή ήταν ενεργή προς τη συνολική χρονική διάρκεια μεταξύ δύο διαδοχικών μετρήσεων. Οι τιμές που προσεγγίζουν τη μονάδα υποδηλώνουν γραμμές που ήταν ενεργές καθ' όλη τη χρονική αυτή διάρκεια.

Εξετάζοντας τα ιστογράμματα, μπορούμε για ακόμα μια φορά να διαπιστώσουμε ότι το *σύνολο 1,* παρουσιάζει διαφορετική συμπεριφορά από τα υπόλοιπα σύνολα, έχοντας συγκεντρωμένες, ως επί το πλείστον, γραμμές με λόγο μικρότερο του ενός δευτέρου, στοιχείο ενδεικτικό της ποιότητας των γραμμών που το απαρτίζουν.

Η αντίθεση σε σχέση με τα προηγούμενα ιστογράμματα που αφορούσαν το *σύνολο 1* είναι η διακύμανση των τιμών γύρω από τη μέση τιμή της μεταβλητής για το συγκεκριμένο σύνολο, όμως το στοιχείο αυτό δεν μπορεί να κριθεί ούτε αρνητικό, ούτε παράδοξο καθώς πολλοί παράγοντες επηρεάζουν το χρόνο κατά τον οποίο είναι ενεργή μια γραμμή. Το σημαντικό δεδομένο που μπορούμε να εξάγουμε από αυτή τη γραμμή είναι η κακή ποιότητα σύνδεσης που παρατηρείται στο σύνολο των γραμμών που απαρτίζουν το *σύνολο 1.*

Αναλυτικό Σετ Δεδομένων 1

	All	Σύνολο 0	Σύνολο 1	Σύνολο 2	Σύνολο 3
Atuc Attainable Rate	8422.95	9206.75	2499.63	9407.25	7686.15
Atuc SNR Margin	10.29	11.22	3.13	10.49	10.84
Atuc dES	82.02	40.96	14.63	2931.81	686.83
Atuc Power	17.29	18.84	5.27	19.15	17.60
Atur Attainable Rate	921.67	1005.79	282.36	977.96	892.34
Atur SNR Margin	10.15	11.07	3.28	9.85	9.72
Atur dES	3.07	1.21	1.30	14.09	91.36
Atur Power	12.48	13.68	3.56	13.15	11.04
Reset	0.06	0.03	0.18	0.18	0.62
dt	3430.97	3780.89	784.84	3616.72	3259.31
Time	0.81	0.88	0.22	0.83	0.78

Αναλυτικό Σετ Δεδομένων 2

	All	Σύνολο 0	Σύνολο 1	Σύνολο 2	Σύνολο 3
Atuc Attainable Rate	8457.41	9151.87	3181.21	7025.66	7880.23
Atuc SNR Margin	10.40	11.27	3.68	8.49	11.62
Atuc dES	61.16	34.83	59.15	333.72	2940.40
Atuc Power	17.31	18.74	6.20	15.38	18.55
Atur Attainable Rate	925.78	1002.49	334.40	828.45	944.81
Atur SNR Margin	10.23	11.08	3.69	10.28	9.51
Atur dES	2.94	1.26	1.86	284.60	6.93
Atur Power	12.62	13.68	4.36	12.34	13.49
Reset	0.11	0.03	0.18	11.06	0.18
dt	3114.97	3413.33	839.59	2382.54	3080.32
Time	0.81	0.88	0.26	0.63	0.81

Πίνακας 5 & 6: Παρουσίαση των μέσων όρων όλων των μεταβλητών. Στη πρώτη στήλη παρατίθεται ο συγκεντρωτικός μέσος όρος, που αφορά το σύνολο των γραμμών, ενώ στις στήλες 2-5 παρατίθενται οι μέσοι όροι που προκύπτουν μόνο για τις γραμμές που ανήκουν στα αντίστοιχα σύνολα.

5.3.3 - Συμπεράσματα

Σε αυτό το εδάφιο θα γίνει η απόπειρα να συνδεθούν όλα τα επιμέρους αποτελέσματα και συμπεράσματα που προέκυψαν από την ανάλυση των ιστογραμμάτων της κάθε μεταβλητής ξεχωριστά, ώστε να καταστεί δυνατή η υποστήριξη του αρχικού μας ισχυρισμού, δηλαδή ότι επετεύχθη ο στόχος που είχε τεθεί για την εύρεση και το διαχωρισμό των γραμμών στις οποίες εμφανίζεται το φαινόμενο του crosstalk.

Όπως αναφέρεται και στο εδάφιο **5.3.1** η έλλειψη ενός συνόλου εκπαίδευσης, αλλά και η μη διαθεσιμότητα κάποιων παραμέτρων που θα μπορούσαν να χαρακτηριστούν καταλυτικές στην ανεύρεση των γραμμών που πάσχουν από crosstalk, η έρευνα επικεντρώθηκε στις παραμέτρους που ήταν διαθέσιμες και εφαρμόζοντας κάποια λογικά κριτήρια για το τρόπο που επηρεάζει το crosstalk αυτές τις παραμέτρους, αποπειράθηκε ο διαχωρισμός αυτών των γραμμών από τις υπόλοιπες.

Σε αυτό το σημείο, μπορεί να ειπωθεί ότι ευνοηθήκαμε και από τον παράγοντα τύχη καθώς, ο αρχικός διαχωρισμός που αποπειραθήκαμε, ο οποίος δεν αποσκοπούσε στην ανεύρεση των γραμμών με ενδείξεις φαινομένου crosstalk, αλλά στον διαχωρισμό των γραμμών σε λειτουργικές και προβληματικές, προέκυψε ένα υποσύνολο γραμμών που όπως φάνηκε από την περαιτέρω ανάλυση περιελάμβανε τη πλειονότητα των γραμμών που συγκεντρώνουν τα χαρακτηριστικά που τα λογικά κριτήρια που είχαμε θέσει υπαγόρευαν ότι χαρακτήριζαν τις γραμμές που πάσχουν από το συγκεκριμένο φαινόμενο.

Πριν γίνει η εξέταση των χαρακτηριστικών των clusters, θεωρήσαμε ότι οι γραμμές που υποφέρουν από το φαινόμενο του crosstalk έχουν χαμηλές τιμές περιθωρίου σηματοθορυβικού λόγου, εξαιτίας του γεγονότος ότι μια αύξηση στα επίπεδα τιμών του θορύβου, που μπορεί να οφείλεται σε παρεμβολές από άλλες πηγές, οδηγεί στη μείωση του **SNR** και συνεπώς μειώνει το περιθώριο σηματοθορυβικού λόγου πάνω από την οριακή τιμή κατά την οποία πετυχαίνουμε πιθανότητα σφάλματος 10^{-7}. Επίσης λογικά έχουν και μειωμένη δυνατότητα επίτευξης υψηλών ρυθμών μετάδοσης. Όταν στο κανάλι επικρατούν συνθήκες οι οποίες δεν επιτρέπουν την εύρυθμη λειτουργία μιας σύνδεσης, όπως η ύπαρξη υψηλών επιπέδων θορύβου, ο μέγιστος επιτεύξιμος ρυθμός μετάδοσης υποχρεωτικά περιορίζεται. Τέλος, γραμμές που πάσχουν από το φαινόμενο του crosstalk, με μεγάλη πιθανότητα, παρουσιάζουν δυσκολίες στο συγχρονισμό και στη παρατεταμένη διατήρηση της σύνδεσης. Έτσι έγινε η επιπλέον υπόθεση ότι οι γραμμές που εμφανίζουν το φαινόμενο, θα χαρακτηρίζονται από περιορισμένο χρόνο ενεργής σύνδεσης.

Η εξέταση των ιστογραμμάτων που σχετίζονται με όλες τις μεταβλητές που περιγράφονται στη προηγούμενη παράγραφο, έδειξε ότι οι λογικές υποθέσεις που είχαν γίνει όχι μόνο ισχύουν, αλλά ότι είχαμε καταφέρει να συγκεντρώσουμε τη πλειονότητα των γραμμών που χαρακτηρίζονται από αυτές τις συνθήκες, σε ένα cluster.

Συγκεκριμένα, οι τιμές στο μέγιστο επιτεύξιμο ρυθμό για το *σύνολο 1*, και για τις δύο κατευθύνσεις του δικτύου, χαρακτηρίζονταν ως ιδιαιτέρως χαμηλές όπως φαίνεται και από τα αντίστοιχα ιστογράμματα αλλά και από τους **πίνακες 5** και **6** στους οποίους είναι συγκεντρωμένες οι μέσες τιμές των μεταβλητών ανά cluster. Επιπλέον, για το ίδιο σύνολο, οι μεταβλητές **mean atuc & atur SNR margin** παρουσιάζουν χαμηλές τιμές, ενδεικτικό της ύπαρξης υψηλών επιπέδων θορύβου στο κανάλι, απόρροια της ύπαρξης ηλεκτρομαγνητικών παρεμβολών από γειτονικές πηγές ηλεκτρομαγνητικών σημάτων. Τέλος, όλες οι γραμμές που ανήκουν στο *σύνολο 1* χαρακτηρίζονται από χαμηλές τιμές της μεταβλητής **time,** η οποία εμπεριέχει το λόγο του ενεργού προς τον συνολικό χρόνο ανάμεσα σε δύο διαδοχικές μετρήσεις από το router του παρόχου.

5.3.4 - Μελλοντικές Κατευθύνσεις

Όπως έχει αναφερθεί σε πολλά σημεία μέσα σε αυτή τη διπλωματική εργασία, η ίδια η φύση του προβλήματος, αλλά και η έλλειψη των εξαιρετικά σημαντικών, σε τέτοιου είδους προβλήματα, δεδομένων εκπαίδευσης, δεν μας δίνουν τη δυνατότητα να είμαστε απόλυτα βέβαιοι για την ορθότητα της κατηγοριοποίησης που πραγματοποιήσαμε, όμως μέσα από αυτή την εργασία θεωρούμε ότι έχει γίνει ένα σημαντικό πρώτο βήμα προς την αυτοματοποιημένη ανεύρεση γραμμών που εμφανίζουν σημάδια ύπαρξης φαινομένου crosstalk, που συμβάλλει στη περαιτέρω προσπάθεια επίλυσης αυτού του ιδιαιτέρως σημαντικού προβλήματος της ADSL κοινότητας.

Μια επέκταση αυτής της εργασίας θα μπορούσε να είναι η προσπάθεια να εφαρμοστούν κάποιες τεχνικές DSM στις γραμμές που έχουν χαρακτηρίσει ως γραμμές που εμφανίζουν crosstalk. Επίσης, θα μπορούσαμε να επανεξετάσουμε και να επαναπροσδιορίσουμε το τρόπο υλοποίησης της παρούσας εργασίας, αν οι παράμετροι που περιγράφονται στο εδάφιο **2.3** καταστούν μελλοντικά διαθέσιμες.

Βιβλιογραφία

[1] ANSI T1.413-1998 "Network and Customer Installation Interfaces – Asymmetric Digital Subscriber Line (ADSL) Metallic Interface", American National Standards Institute, 1998

[2] ITU-T G.992.1 "Asymmetric Digital Subscriber Line (ADSL) Transceivers", International Telecommunication Union, 1999

[3] ITU-T G.992.5 "Asymmetric Digital Subscriber Line (ADSL) Transceivers - Extended Bandwidth ADSL2 (ADSL2+)", International Telecommunication Union, 2005

[4] Philip Golden, Hervé Dedieu, and Krista S. Jacobsen, "Implementation And Applications Of DSL Technology", Boca Raton, FL: Auerbach Publications, 2008

[5] Philip Golden, Hervé Dedieu, and Krista S. Jacobsen, "Fundamentals Of DSL Technology", Boca Raton, FL: Auerbach Publications, 2006

[6] "Dynamic Spectrum Management", < http://en.wikipedia.org/wiki/Dynamic_Spectrum_Management>

[7] J. A. Hartigan, "Clustering Algorithms", New York: Wiley, 1975

[8] Richard O. Duda, Peter E. Hart, David G. Stork, "Pattern Classification (2nd Edition)" New York: Wiley, 2001

[9] N. Metropolis, S. Ulam, "The Monte Carlo Method", *Journal of the American Statistical Association* (American Statistical Association), 1949