ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Εποπτεία συσκευών για την αναγνώριση έκτοπων τιμών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Αναγνωστίδη Σωτήριου Κωνσταντίνου

Επιβλέπων: Ανδρέας Σταφυλοπάτης

Καθηγητής Ε.Μ.Π

Αθήνα, Ιούλιος 2018

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Εποπτεία συσκευών για την αναγνώριση έκτοπων τιμών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Αναγνωστίδη Σωτήριου Κωνσταντίνου

Επιβλέπων: Ανδρέας Σταφυλοπάτης
Καθηγητής Ε.Μ.Π

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή της 18ης Ιουλίου 2018.

(Υπογραφή)	(Υπογραφή)	(Υπογραφή)
.....	.....	.....
Ανδρέας-Γεώργιος Σταφυλοπάτης	Παναγιώτης Τσανάκας	Γιώργος Στάμου
Καθηγητής Ε.Μ.Π	Καθηγητής Ε.Μ.Π	Αναπληρωτής Καθηγητής Ε.Μ.Π

Αθήνα, Ιούλιος 2018

(Υπογραφή)

.....

Αναγνωστίδης Σωτήριος Κωνσταντίνος

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π

Copyright ©Αναγνωστίδης Σωτήριος Κωνσταντίνος, 2018. Με επιφύλαξη παντός δικαιώματος. All rights reserved Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς το συγγραφέα. Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευτεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσοβίου Πολυτεχνείου.

Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε στο Εργαστήριο Ευφυών Συστημάτων της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου, υπό την επίβλεψη του Καθηγητή Ανδρέα-Γεωργίου Σταφυλοπάτη.

Θα ήθελα να ευχαριστήσω τον καθηγητή μου κ. Ανδρέα-Γεώργιο Σταφυλοπάτη, για την εποπτεία κατά την εκπόνηση της εργασίας μου, τις γνώσεις που μου προσέφερε με τη διδασκαλία, καθώς και την εμπιστοσύνη που μου έδειξε καθ΄ όλη τη διάρκεια της διαδικασίας.

Θα ήθελα επίσης να ευχαριστήσω για τη συνεργασία μας και τις πολύτιμες συμβουλές του τον κ. Χριστόφορο Αναγνωστόπουλο, Λέκτορα του Imperial College London. Η καθοδήγηση του κατά τη διάρκεια εκπόνησης αυτής, η ενθάρρυνσή του και η υπομονή του ήταν πολύτιμες, ενώ χωρίς τη συμβολή του η ολοκλήρωση αυτής της εργασίας δε θα ήταν εφικτή.

Αναγνωστίδης Σωτήριος-Κωνσταντίνος

Περίληψη

Με τον αριθμό των διασυνδεδεμένων στο διαδίκτυο συσκευών - χρηστών να αυξάνεται συνεχώς, υπάρχει μια ολοένα αυξανόμενη ανάγκη στενής παρακολούθησης αυτών, καθώς και των ροών που αυτές παράγουν. Η παρακολούθηση αυτή οφείλει να λαμβάνει χώρα σε πραγματικό χρόνο και να προσαρμόζεται στις δυναμικές συμπεριφορές των χρηστών.

Στην παρούσα διπλωματική εργασία παρουσιάζουμε ένα μηχανισμό κατηγοριοποίησης των ανά συσκευή ροών αυτών, σε ορισμένες δυναμικές καταστάσεις. Η δαδικασία αποσκοπεί στη σημαντική μείωση της πληροφορίας. Στη συνέχεια μοντελοποιούμε τις μεταβάσεις των συσκευών μεταξύ των καταστάσεων αυτών. Τέλος ομαδοποιούμε τους χρήστες αυτούς σύμφωνα με τις μεταβάσεις όπως αυτές διαμορφώθηκαν, με στόχο τη δημιουργία ορισμένων προφίλ συμπεριφοράς.

Η παραπάνω διαδικασία μας επιτρέπει να καταλήξουμε σε χρήσιμες πληροφορίες σχετικά με την κίνηση των συσκευών, ενώ παράλληλα οδηγεί και στην απομόνωση συσκευών με ιδιάζουσα συμπεριφορά. Επιτρέπει με τον τρόπο αυτό σε ένα διαχειριστή συστήματος τον περιορισμό των χρηστών που αξίζουν περαιτέρω διερεύνηση.

Λέξεις Κλειδιά: Ανίχνευση Ανωμαλιών, αλγόριθμος Expectation Maximization πραγματικού χρόνου, απόσταση Kullback - Leibler, Κρυφό Μαρκοβιανό Μοντέλο

Abstract

While the number of online devices - users continues to rise, there is an increasing need for close monitoring of these devices, as well as the flows they generate. This monitoring should take place in a real time context and be adapted to the dynamic behaviors of these users.

In this diploma thesis we present a mechanism for categorizing the device flows in a number of dynamic states. This process is aimed at significantly reducing the volume of useful information. Then we model the transitions of the devices between these states. Finally, we group these users according to the transitions that they produced in order to identify certain behavioral profiles.

The above procedure allows us to come up with useful information about the flows generated by the devices, while also leading to the isolation of devices with irregular behavior. This system can greatly assist administrators by restricting users they should pay more attention to.

Keywords: online Expectation Maximization, Kullback - Leibler divergence, Hidden Markov Model, Anomaly Detection

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

1.1 Αντικείμενο

Η συνεχής ανάπτυξη του διαδικτύου και ο συνεχής πολλαπλασιασμός των συσκευών που είναι διαρκώς συνδεδεμένες στο διαδίκτυο ή οποιασδήποτε άλλης μορφής εταιρικού δικτύου είναι γεγονός. Η ίδια τάση επιβεβαιώνεται από καθημερινές απλές συσκευές, που ακολουθώντας την ίδια αυτή προδιάθεση, συνδέονται τόσο μεταξύ τους όσο και με το internet, στα πλαίσια του Internet of Things.

Η κλιμακούμενη αυτή τάση έχει οδηγήσει και σε μια πρωτοφανή αντίστοιχη αύξηση της κίνησης των δικτύων. Το γεγονός αυτό καθιστά πολύ δύσκολο τον έλεγχο αυτής της κίνησης και τον εντοπισμό ανωμαλιών σε αυτήν. Ειδικότερα, σε μεγάλους οργανισμούς, που έρχονται αντιμέτωποι με εκατομμύρια συνδέσεις κάθε μέρα, είναι αναγκαία η ύπαρξη κατάλληλου μηχανισμού και συστήματος έλεγχου αυτής. Υπάρχοντα συστήματα πολλές φορές αδυνατούν να ανταποκριθούν στις επικρατούσες συνθήκες. Πολλά συστήματα του παρελθόντος βασίζονται σε συγκεκριμένους κανόνες, ορισμένους από τους διαχειριστές του συστήματος. Οι κανόνες αυτοί ωστόσο, αν και αποτελούν αποτελεσματικές λύσεις κατά περιόδους, αδυνατούν να περιγράψουν το ευρύτερο σύνολο των πιθανών περιπτώσεων, ενώ δεν καταφέρνουν να προσαρμοστούν σε πιθανές νέες ή δυναμικές συμπεριφορές. Ενδεχόμενες αλλαγές της αποδεκτής συμπεριφοράς απαιτούν αρκετή προσωπική εργασία από την πλευρά των διαχειριστών, με στόχο την αναπροσαρμογή των εκάστοτε κανόνων.

Στην παρούσα διπλωματική παρουσιάζουμε έναν αλγόριθμο ομαδοποίησης της κίνησης ενός συνόλου χρηστών. Στη συνέχεια μοντελοποιούμε τις μεταβάσεις των χρηστών μεταξύ των διαφορετικών ομάδων (clusters) που έχουν δημιουργηθεί. Σκοπός είναι η δημιουργία προφίλ συμπεριφοράς και η κατάταξη των χρηστών σε κάποιο εξ αυτών. Βασική προϋπόθεση του αλγορίθμου είναι η δυνατότητα επεξεργασίας

δεδομένων σε πραγματικό χρόνο όπως καταφθάνουν αυτά, καθώς και η δυνατότητα προσαρμογής σε αλλαγές στα δεδομένα αυτά.

1.2 Δομή διπλωματικής

Η παρούσα διπλωματική αποτελείται στο σύνολο από έξι κεφάλαια. Η δομή αυτής διπλωματικής έχει ως εξής:

Στο δεύτερο κεφάλαιο παρουσιάζουμε και αναλύουμε ορισμένες θεωρητικές ιδέες που θα μας χρησιμεύσουν στη συνέχεια. Αυτές αφορούν ορισμένες τεχνικές που έχουν εφαρμοστεί ήδη στο χώρο της ανίχνευσης ανωμαλιών και της ομαδοποίησης (clustering), καθώς και βασικά μοντέλα και θεωρητικές έννοιες, των οποίων θα γίνει χρήση στη συνέχεια.

Στο τρίτο κεφάλαιο αναλύουμε με μεγαλύτερη σαφήνεια και ακρίβεια το πρόβλημα που θέλουμε να αντιμετωπίσουμε. Παραθέτουμε απλές τεχνικές και πιθανά μειονεκτήματα που μπορούν να προκύψουν από αυτές.

Στο τέταρτο κεφάλαιο αναλύουμε διαφορετικές προσεγγίσεις για το πρόβλημα, πλεονεκτήματα και ελαττώματα αυτών. Τελικά οδηγούμαστε στο τελικό μοντέλο που θα χρησιμοποιήσουμε.

Στο πέμπτο κεφάλαιο ελέγχουμε την απόδοση του μοντέλου σε σχέσεις με διαφορετικές τεχνικές. Για το σκοπό αυτό χρησιμοποιούμε επισημασμένα, τεχνητά δημιουργημένα δεδομένα, καθώς και πραγματικά δεδομένα κίνησης χρηστών.

Τέλος, στο έκτο κεφάλαιο ανακεφαλαιώνουμε και παρουσιάζουμε ορισμένες περιοχές περαιτέρω βελτίωσης, που παρουσιάζουν ιδιαίτερο ενδιαφέρον για μελλοντικές εργασίες.

ΚΕΦΑΛΑΙΟ 2

Θεωρητικό Υπόβαθρο

2.1 Ανίχνευση Ανωμαλιών

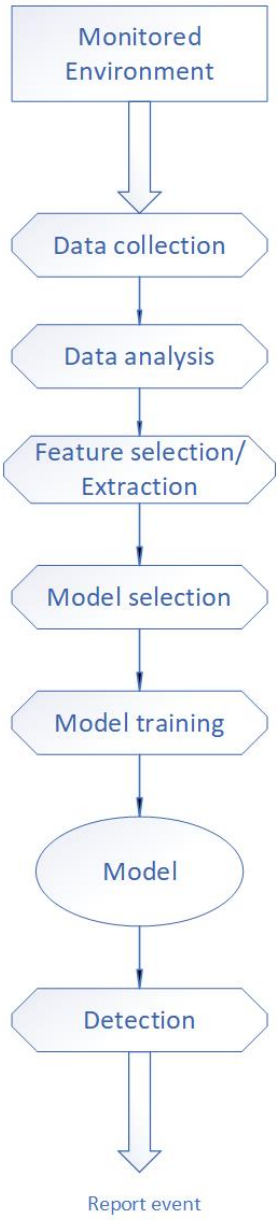
Με τον όρο Ανίχνευση Ανωμαλιών αναφερόμαστε στην εύρεση μεμονωμένων συμβάντων ή και ολόκληρων ακουλιθιών, τα οποία δεν συμβαδίζουν με την αναμενόμενη συμπεριφορά ενός συστήματος. Η Ανίχνευση Ανωμαλιών αποτελεί αναπόσπαστο κομμάτι μιας πληθώρας εφαρμογών. Παραδείγματα τέτοιων εφαρμογών αποτελούν απόπειρες απάτης σε συναλλαγές πιστωτικών καρτών, δικτυακές επιθέσεις σε ένα δίκτυο υπολογιστών ή η αναγνώριση έκτπων τιμών σε θέματα υγείας. Η εφαρμογή Τεχνητής Νοημοσύνης σε θέματα υγείας αποτελεί κλάδο με ιδιαίτερο ενδιαφέρον και περιθώρια ανάπτυξης.

Η σημασία εύρεσης των ανωμαλιών αυτών έγκειται στο γεγονός ότι αυτές συνήθως αποτελούν ενδιαφέρουσες πληροφορίες, κρίσιμες για το σύστημα και τον διαχειριστή αυτού. Για παράδειγμα, περίοδος αυξημένης κίνησης στα πλαίσια ενός δικτύου, μπορεί να οφείλεται σε απόπειρα κακόβουλης απόπειρας διείσδυσης, ενώ σε εικόνες MIR μπορεί να οδηγήσει στην εύρεση επικίνδυνων όγκων ή καρκινικών κυττάρων. Τα τελευταία χρόνια στο συγκεκριμένο κλάδι, παρατηρείται ένα αυξανόμενο ενδιαφέρον με την ανάπτυξη ποικίλων τεχνικών, μερικές εκ των οποίων παρατίθενται και στην παρούσα εργασία.

Σε σχέση με τα υπόλοιπα πεδία της Μηχανικής Μάθησης, παρουσιάζονται ορισμένες προκλήσεις. Κύρια δυσκολία έγκειται στον ορισμό των περιοχών, η μετάβαση εκατέρωθεν των οποίων σηματοδοτεί τη διαφοροποίηση μεταξύ ανωμαλιών και ομαλών συμβάντων. Επίσης όπως προαναφέρθηκε, η συνεχής εξέλιξη της τεχνολογίας καθιστά απαραίτητη τη δυνατότητα δυναμικής αναπροσαρμογής, όσον αφορά τον τρόπο με τον οποίο ορίζεται η αποδεκτή συμπεριφορά. Σημαντική πρόκληση αποτελεί επιπλέον η μεγάλη ανισορροπία σχετικά με το πλήθος το δεδομένων στις δύο

κατηγορίες ανωμαλιών ή μη. Το μικρό πλήθος δεδομένων στην πρώτη κατηγορία αυτομάτως αποκλείει την αυτούσια εφαρμογή πολλών τεχνικών κατηγοριοποίησης (classification), τεχνικές που παρουσιάζουν ιδιαίτερα αξιοσημείωτα αποτελέσματα σε διάφορες περιπτώσεις. Τέλος, η γνώση της δομής υπαρχόντων συστημάτων ανίχνευσης επιτρέπει στους εισβολείς τη λήψη κατάλληλων μέσων για την απόκρυψη όποιας μετέπειτας κακόβουλης πράξης.

Κατά κανόνα, η διαδικασία ανίχνευσης αποτελείται από τα στάδια που φαίνονται στο Σχήμα 2.1.



Σχήμα 2.1. Διαδικασία ανίχνευσης

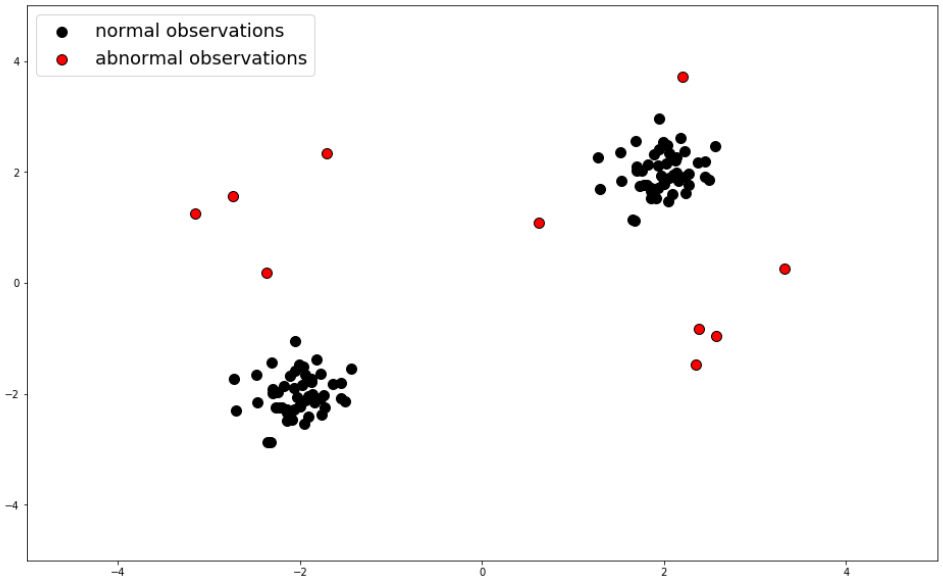
Κατά τη διάρκεια της εκπαίδευσης (training), το μοντέλο χτίζεται με βάση αποδεκτές συμπεριφορές. Κατά τη διάρκεια της ανίχνευσης (detection), αν βρεθεί συμπεριφορά εκτός της αποδεκτής, αυτή αναφέρεται. Προϋπόθεση πολλές περιπτώσεις είναι το να αποτελεί μια συνεχή διαδικασία κατά την οποία το μοντέλο που έχουν επιλέξει θα πρέπει να ανταποκρίνεται στις αλλαγές του περιβάλλοντός του. Κατ΄ αυτήν την έννοια το στάδιο της ανίχνευσης στις περιπτώσεις αυτές θα πρέπει να συνδέεται και με το στάδιο της εκπαίδευσης, καθώς νέα δεδομένα επιφέρουν αλλαγές στα χαρακτηριστικά του μοντέλου.

2.2 Τύποι Ανωμαλιών

Σε αυτό το σημείο θα ήταν σημαντικό να επισημάνουμε τα διαφορετικά είδη ανωμαλιών που εμφανίζονται.

1. Σημειακές ανωμαλίες.

   Ο πιο απλός τύπος ανωμαλιών εμφανίζεται όταν ένα μεμονωμένο σημείο εμφανίζεται ως ανωμαλία σε σχέση με τα υπόλοιπα σημεία που διαθέτουμε. Παράδειγμα σημειακής ανωμαλίας φαίνεται στο Σχήμα 2.2.



Σχήμα 2.2. Σημειακές ανωμαλίες

2. Ανωμαλίες ανάλογα με το περιβάλλον (contextual anomalies).

   Σε αυτή την κατηγορία ανήκουν σημεία, τα οποία αναλόγως με την οπτική και τον τρόπο αντιμετώπισης τους μπορεί να αποτελούν ανωμαλίες ή όχι. Για

παράδειγμα, μια υψηλή τιμή βροχόπτωσης μπορεί να αποτελεί ανωμαλία κατά τους θερινούς μήνες, αλλά να θεωρείται φυσιολογική όταν πρόκειται για χειμερινό μήνα. Ομοίως, ο αριθμός παλμών της καρδίας ενός ανθρώπου θα πρέπει να συνοδεύεται πάντοτε από τις συνθήκες κάτω από τις οποίες αυτός μετρήθηκε. Αντιστοίχως, υψηλή κίνηση σε κάποιο δίκτυο, αισθητά μεγαλύτερη από τον μέσο όρο, μπορεί να δικαιολογηθεί λαμβάνοντας υπόψη το χρονικό πλαίσιο κατά το οποίο παρατηρήθηκε το συμβάν.

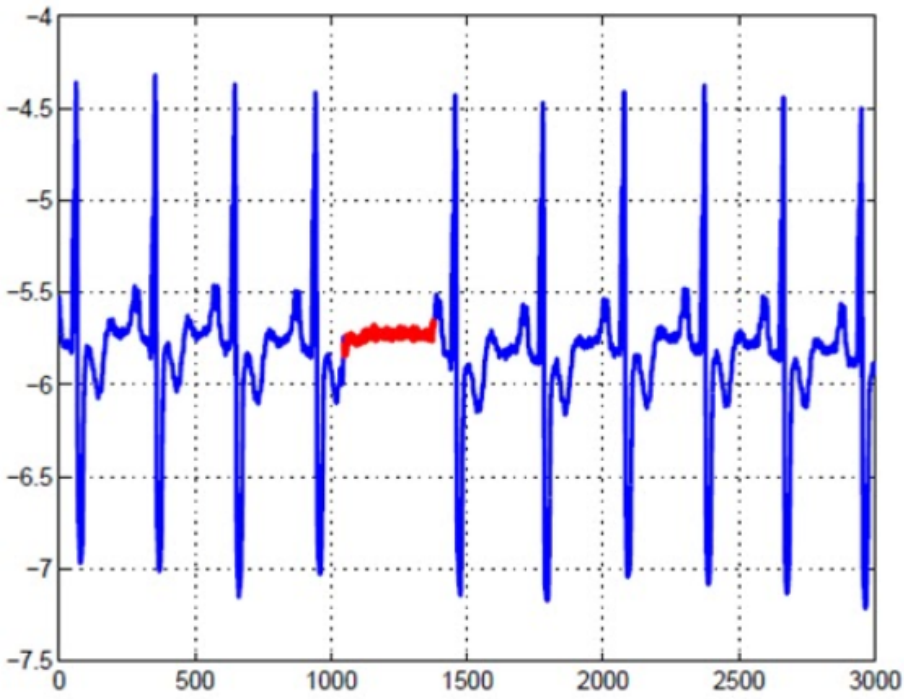
Ο ορισμός του κατάλληλου πλαισίου, με βάση το οποίο θα αξιολογηθεί ένα συμβάν και θα ταξινομηθεί ως ανωμαλία ή όχι, είναι υψίστης σημασίας και αποτελεί μία από τις σημαντικότερες και δυσκολότερες προκλήσεις του κλάδου.

3. Συλλογικές ανωμαλίες (collective anomalies)

Ανωμαλίες αυτής της κατηγορίας προκαλούνται όχι από ένα μεμονωμένο συμβάν, αλλά από συνδυασμό συμβάντων. Σε συνέχεια του προηγούμενου παραδείγματος, ελαφριά βροχόπτωση έναν εαρινό μήνα μπορεί να μην αποτελεί από μόνο του ανωμαλία. Η επανάληψη του ιδίου συμβάντος ωστόσο για πολλές μέρες συνεχόμενες, μπορεί να σηματοδοτηθεί ως ανωμαλία. Χαρακτηριστικό παράδειγμα αποτελεί επίσης η έξοδος ενός ηλεκτροκαρδιογραφήματος. Κάποια χαμηλή έξοδος δεν αποτελεί ανωμαλία από μόνη της, αλλά ακολουθία διαδοχικών εξόδων ορίζει το τελικό αποτέλεσμα, όπως παρατηρούμε και στο Σχήμα 2.3.

2.3 Μοντέλα

Οι περισσότεροι εκ των αλγορίθμων που πραγματοποιούν Ανίχνευση Ανωμαλιών, μερικοί εκ των οποίων θα παρουσιαστούν στη συνέχεια, βασίζονται στην ύπαρξη κάποιου μοντέλου. Η γνώση αυτού επήλθε κατά τη διαδικασία της εκπαίδευσης ή/και γνώσεις και εμπειρίες επαγγελματιών στο εκάστοτε πεδίο. Ρόλος του μοντέλου είναι η καλύτερη δυνατή περιγραφή της κατανομής των δεδομένων που διαθέτουμε. Το μοντέλο αυτό καθορίζει και την τελική απόφαση για τα νέα δεδομένα που καταφθάνουν. Τα χαρακτηριστικά του μοντέλου μπορεί να είναι καθορισμένα εξ αρχής από το διαχειριστή, όπως το πλήθος των κατανομών που θα επιλεγεί για την αναπαράσταση και το είδος των κατανομών αυτών [26], ή μπορεί να μεταβάλλονται κατά τη διάρκεια, με στόχο την καλύτερη αναπαράσταση νέων δεδομένων [7].



Σχήμα 2.3. Συλλογικές ανωμαλίες

2.4 Κατηγορίες αλγορίθμων ανίχνευσης ανωμαλιών

Οι περισσότεροι αλγόριθμοι που έχουν αναπτυχθεί, ανήκουν σε κάποια από τις κατώτερες κατηγορίες:

- κατηγοριοποίηση (classification)
- ομαδοποίηση (clustering)
- κοντινότερου γείτονα (nearest neighbor)
- στατιστικές μέθοδοι (statistical)

Στις συνέχεια θα παρουσιασθούν κάποιες από τις πιο βασικές μεθόδους που έχουν αναπτυχθεί, ορισμένες ιδέες από τις οποίες θα εφαρμοστούν και στην παρούσα εργασία.

2.4.1 Κατηγοριοποίηση

Χρησιμοποιείται ένα μοντέλο, η εκπαίδευση του οποίου πραγματοποιείται σε κάποια δεδομένα (train set) που ανήκουν σε ένα σύνολο διαφορετικών κλάσεων, ενώ στη συνέχεια κατηγοριοποιεί νέα δεδομένα σε κάποια από τις διαφορετικές κλάσεις που

ήδη γνωρίζει. Μπορούμε να διακρίνουμε δύο διαφορετικές περιπτώσεις, ανάλογα με τον τρόπο λειτουργίας του.

Μπορούν να δίνονται σαν δεδομένα εκπαίδευσης τιμές από όλες τις κατηγορίες. Αυτό θα σημαίνει τη διάθεση δεδομένων τόσο από φυσιολογικές συμπεριφορές όσο και από ύποπτες. Η πρακτική αυτή ωστόσο, εμφανίζει κάποια προφανή μειονεκτήματα. Σε πραγματικές εφαρμογές οι περιπτώσεις ανωμαλιών είναι σπάνιες. Το γεγονός αυτό οφείλεται τόσο στο ότι οι περιπτώσεις παραβιάσεων συμβαίνουν σπάνια στην πράξη, όσο και στο ότι ακόμα λιγότερες είναι οι φορές που μπορούν να αναλυφθούν και να εξακριβωθεί ότι πρόκειται όντως για περίπτωση που θέλουμε να αποφύγουμε στο μέλλον. Επομένως, οι κλάσεις θα είναι υπερβολικά αντιζυγισμένες, οδηγώντας με μια απλή υλοποίηση σε πολλές εσφαλμένες ανιχνεύσεις (false negatives). Σημαντικό μειονέκτημα επιπλέον, αποτελεί το ότι οι ανωμαλίες παρουσιάζουν μεταξύ τους μια ποικιλότητα. Η ποικιλία αυτή σημαίνει δύσκολο προσδορισμό του μοντέλου που θα τις περιγράφει. Παράλληλα, νέου είδους ανωμαλίες δημιουργούνται με στόχο να εκμεταλλευτούν τέτοιου είδους αδυναμίες συστημάτων.

Στη δεύτερη κατηγορία δίνονται δεδομένα από την κατηγορία φυσιολογικών συμπεριφορών. Επομένως, η ανάδειξη ανωμαλιών ανατίθεται στον προσδιορισμό της ομοιότητας ενός νέου δεδομένου, σε σχέση με τα υπάρχοντα. Εφόσον εμπίπτει σε κάποια πλαίσια, γίνεται αποδεκτή, ενώ σε αντίθετη περίπτωση, όχι. Η μέθοδος αυτή μπορεί να εμπλουτιστεί με ένα σύστημα επιβράβευσης ή τιμωρίας από τους διαχειριστές, εφόσον έχει ληφθεί σωστή ή λανθασμένη απόφαση αντίστοιχα.

Βασικές μέθοδοι που έχουν αναπτυχθεί είναι τα Νευρωνικά Δίκτυα, Bayessian Models, SVM, Συστήματα Κανόνων. Τα τελευταία εμφανίζουν ιδιαίτερη άνθιση λόγω του μικρού υπολογιστικού κόστους που απαιτούν [37, 40].

2.4.2 Ομαδοποίηση

ΟΙ τεχνικές ομαδοποίησης βασίζονται στην υπόθεση ότι οι φυσιολογικές συμπεριφορές εμφανίζονται πιο συχνά και παρουσιάζουν μεταξύ τους μεγάλες ομοιότητες. Με αυτόν τον τρόπο μπορούν να ομαδοποιηθούν σε αντίθεση με τις ανωμαλίες. Μια διαφορετική ιδέα βασίζεται στο γεγονός, ότι οι ανωμαλίες θα βρίσκονται πιο μακριά από τα υπάρχοντα κέντρα των ομάδων.

Σε αυτές τις περιπτώσεις το στάδιο της ανίχνευσης επιτυγχάνεται συγκρίνοντας την πυκνότητα των ομάδων που έχουν δημιουργηθεί, καθώς και τις αποστάσεις μεταξύ αυτών στο χώρο των χαρακτηριστικών. Για την επίτευξη δυναμικής συμπεριφοράς, οι πυκνότητες αυτές μπορούν να φθίνουν με το χρόνο για την αναπροσαρμογή σε νέα δεδομένα [7]. Εφαρμογές βρίσκουν αλγόριθμοι που βασίζονται στα SOM

[19], K-Means, K-Medoids, EM CLustering. Στην πρώτη περίπτωση, αναθέτουμε τα δεδομένα σε k το πλήθος ομάδες (παράμετρος η οποία αποτελεί και υπερπαράμετρο του συστήματός μας). Η διαφορά του K-Medoids έγκειται στο ότι κάθε ομάδα δεδομένων (cluster) αναπαρίσταται από τη διάμεσο των δεδομένων. Με τον τρόπο αυτό, το στοιχείο που αναπαριστά κάθε ομάδα είναι ανεξάρτητο από ανωμαλίες, σε μεγαλύτερο βαθμό, βελτιώνοντας έτσι τα τελικά αποτελέσματα της ποιότητας των ομάδων, σε ορισμένες περιπτώσεις [36] . Χρησιμοποιώντας EM Clutering, σε αντίθεση με τις προηγούμενες περιπτώσεις, κάθε δεδομένο δεν ανατίθεται σε μοναδική ομάδα. Δεν υπάρχουν αυστηρά κριτήρια διαχωρισμού, αλλά ανατίθενται ποσοστά συμμετοχής σε κάθε ομάδα.

Σημαντικό μειονέκτημα των μεθόδων αυτών αποτελεί το γεγονός, ότι, αν οι ανωμαλίες δημιουργούν μεταξύ τους πυκνές ομάδες επαρκούς μεγέθους, οι τεχνικές αυτές αποτυγχάνουν.

2.4.3 Τεχνικές κοντινότερου γείτονα

Σημαντική τεχνική αποτελούν οι αλγόριθμοι που βασίζονται σε τοπικές ανωμαλίες (Local Outliers). Ορίζεται μια περιοχή ενδιαφέροντος για κάθε δεδομένο και ανάλογα με τη πυκνότητα στην περιοχή αυτή, καθώς και την πυκνότητα αντίστοιχων περιοχών γειτόνων του, εντοπίζονται ανωμαλίες. Η επιλογή της περιοχής αυτής συνήθως ορίζεται με βάση τον k^{th} κοντινότερο γείτονα. Σύμφωνα με αυτόν, ανατίθεται ένας δείκτης σε κάθε δεδομένο, που αποτελεί τον δείκτη ανίχνευσης αυτού [6] .

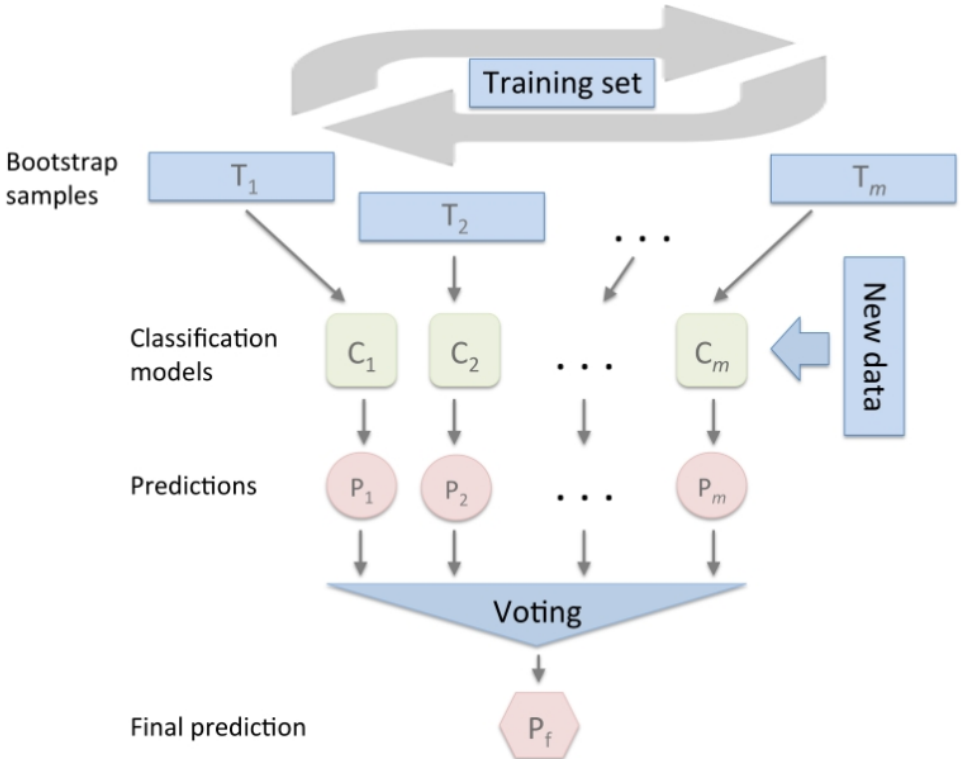
Εμπόδιο των μεθόδων αυτών αποτελεί το συνήθως μεγάλο υπολογιστικό κόστος που απαιτείται για την ολοκλήρωσή τους, καθώς και τα υψηλά ποσοστά false positive που μπορούν να παρατηρηθούν, εφόσον δεν διαθέτουμε επαρκείς περιπτώσεις με παρόμοια χαρακτηριστικά.

2.4.4 Στατιστικές μέθοδοι

Αποδεκτές συμπεριφορές εμφανίζονται σε περιοχές του μοντέλου με μεγαλύτερη πυκνότητα πιθανότητας, ενώ το αντίθετο συμβαίνει με τις ανωμαλίες. Η προσοχή σε αυτήν την περίπτωση εναποτίθεται στον σωστό ορισμό του μοντέλου. Κλασική περίπτωση χρήσης αποτελεί ο ορισμός ενός γκαουσιανού μοντέλου ή ακόμα και ο συνδυασμός περισσότερων για την πιο ακριβή αναπαράσταση των δεδομένων.

Σαφώς υπάρχουν τεχνικές που συνδυάζουν περισσότερες από μία κατηγορίες. Μια ενδιαφέρουσα εφαρμογή είναι ο συνδυασμός πολλαπλών πολύ απλών μεθόδων στο σύνολό τους και η λήψη απόφασης με βάση την πλειοψηφία των αποτεμεσμάτων

αυτών, όπως βλέπουμε και στο Σχήμα 2.4.



Σχήμα 2.4. Αλγόριθμος πλειοψηφίας

Από τα παραπάνω μπορούμε να συμπεράνουμε την πληθώρα των διαφορετικών μεθόδων που υπάρχουν. Η σωστή επιλογή κατάλληλης τεχνικής ή η χρήση υπαρχόντων για τη δημιουργία μίας νέας θα πρέπει να λαμβάνει χώρα με βασικό γνώμονα το είδος του προβλήματος που αντιμετωπίζεται. Η διαφορετική φύση ενός προβλήματος μπορεί να οδηγήσει σε εσφαλμένα συμπεράσματα ενός φαινομενικά επιτυχημένου αλγορίθμου. Ο τύπος των ανωμαλιών (outlier, inlier, etc), θα πρέπει να αναλυθεί εκτενώς εξ αρχής.

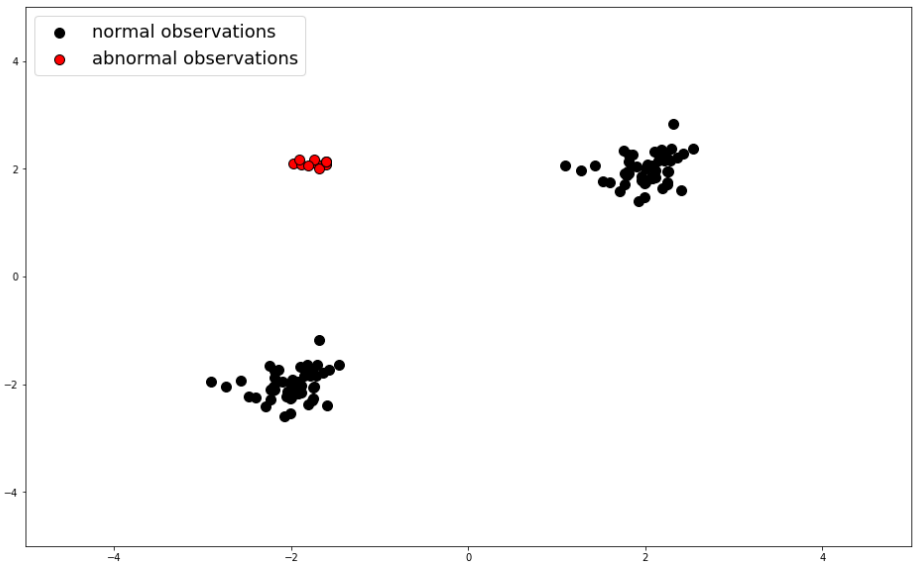
Ο χώρος των διαστάσεων έχει επίσης σημαντικό ρόλο. Τεχνικές μείωσης διαστάσεων μπορούν να εφαρμοστούν. Ωστόσο, αυτές λειτουργούν υπό την παραδοχή ότι τα συμπεράσματά που θα προκύψουν θα είναι ασφαλή και άρτια, ύστερα από την προβολή των διανυσμάτων στο χώρο χαμηλότερης διάστασης που θα προκύψει.

Τέλος, σημαντικός παράγοντας είναι και η συχνότητα των ανωμαλιών. Οι περισσότερες τεχνικές που αναλύθηκαν λειτουργούν με βάση το ότι οι ανωμαλίες εμφανίζονται σποραδικά και με ακανόνιστο τρόπο μεταξύ τους. Ωστόσο, σε πολλές περιπτώσεις, ιδιαίτερα σε δικτυακές εφαρμογές, το ποσοστό των μηνυμάτων που παρουσιάζουν κακοήθεις βλέψεις είναι αξιοπρόσεκτο. Ένας τρόπος αντιμετώπισης του προβλήματος αυτού είναι η ομαδοποίηση τέτοιων μηνυμάτων και η εμφάνιση τους ως

μια καθολική περίσταση με μεγαλύτερη ενδεχομένως βαρύτητα.

2.5 Πολλαπλές ανωμαλίες

Μέχρι αυτό το σημείο αναφερθήκαμε σε μοναδικές ανωμαλίες, είτε αυτές αποτελούνται από ένα σημείο ή από μια ακολουθία σημείων. Η ανίχνευσή τους καθίσταται αρκετά πιο περίπλοκη διαδικασία, όταν ένα πλήθος ανωμαλιών εμφανίζει παρόμοια χαρακτηριστικά μεταξύ τους, όπως παρατηρούμε στο Σχήμα 2.5. Πολλοί αλγόριθμοι που βασίζονται στην σπανιότητα ανωμαλιών για να πάρουν μια απόφαση ενδεχομένως καταλήγουν σε εσφαλμένα αποτελέσματα.



Σχήμα 2.5. Πολλαπλές ανωμαλίες με παρόμοια χαρακτηριστικά

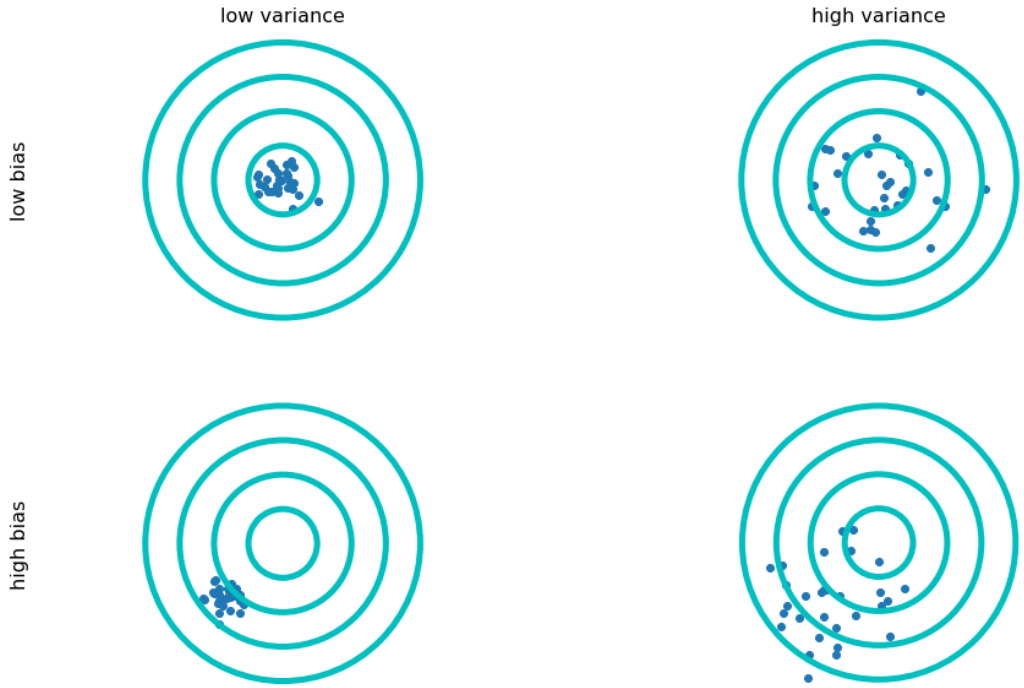
Συχνά ωστόσο θα θέλαμε το μοντέλο να αναγνωρίζει δυναμική συμπεριφορά και να προσαρμόζεται σε νέα δεδομένα. Πυκνός χώρος σημείων που δεν έχουν εμφανιστεί στο παρελθόν, μπορεί να υποδηλώνει νέα αποδεκτή συμπεριφορά στο μέλλον. Παρατηρούμε, επομένως σχέση ανταγωνισμού ανάμεσα στις δύο βασικές αυτές έννοιες: την προσαρμογή σε νέα δεδομένα και την αναγνώριση ανωμαλιών.

2.6 Συμβιβασμός ποικιλότητας - προκατάληψης

Ένα από τα μεγαλύτερα ζητήματα που θα μας απασχολήσει στη συνέχεια είναι ο συμβιβασμός μεταξύ αποδοχής μεγαλύτερης ποικιλότητας στα δεδομένα, ή ο ορισμός προκαθορισμένων όριων σε αυτά (variance - bias tradeoff). Μοντέλα με

μεγάλη προκατάληψη μπορούν συνήθως να γενικευτούν με καλύτερα αποτελέσματα, ωστόσο δεν επεξηγούν με ακρίβεια τα δεδομένα αυτά (underfitting). Από την άλλη πλευρά, μοντέλα με μεγαλύτερη ποικιλότητα παρουσιάζουν με καλύτερη ακρίβεια τα δεδομένα. Ωστόσο εγκυμονεί ο κίνδυνος της υπερεκπαίδευσης και υπερεξειδίκευσης υπό τα δεδομένα αυτά (overfitting).

Τα λάθη στις προβλέψεις σε νέα δεδομένα που καταφθάνουν, μπορούν να γίνουν αντιληπτά ως ένα άθροισμα λαθών ποικιλότητας και προκατάληψης, καθώς και ενός όρου θορύβου των δεδομένων [17]. Αν συμβολίσουμε με Y τα δεδομένα που θέλουμε να προβλέψουμε και X αυτά που γνωρίζουμε, ενώ υποθέσουμε μια σχέση της μορφής $Y = f(X) + e$ τότε λαμβάνουμε: $Err(x) = Bias^2 + Variance + IrreducibleError$ Στο Σχήμα 2.6 παρατηρούμε διαφορετικούς συνδυασμούς των δύο αυτών δεικτών που μας απασχολούν:

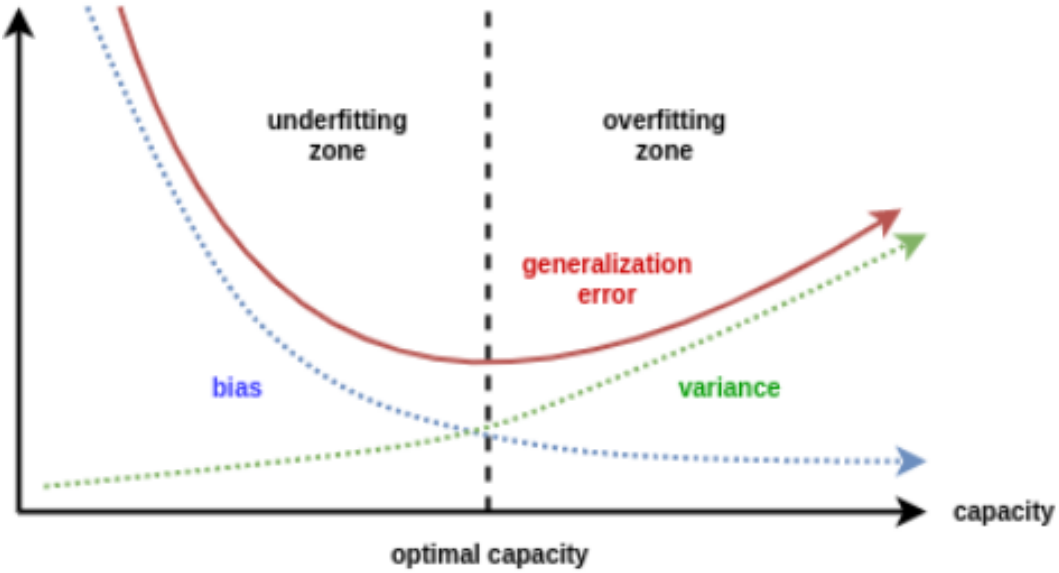


Σχήμα 2.6. Ποικιλότητα και προκατάληψη (variance - bias)

Η κατάλληλη επιλογή του μοντέλου που συνδυάζει καλύτερα τους δείκτες αυτούς, αποτελεί κρίσιμη απόφαση. Αυξάνοντας τον αριθμό των παραμέτρων, αυξάνουμε και την ποικιλότητα, δηλαδή το σύνολο των διαφορετικών δεδομένων που μπορούν να αντιπροσωπευτούν με ακρίβεια, ωστόσο μειώνεται όποια προκατάληψη.

Στόχος είναι η εύρεση του σημείου κατά το οποίο η αύξηση της προκατάληψης συνεπάγεται την μείωση της πολυπλοκότητας και αντίστροφα. Αν υπερβούμε το σημείο αυτό, καταφεύγουμε σε overfitting, ενώ σε διαφορετική περίπτωση, σε underfitting

(Σχήμα 2.7). Στην πράξη, συνήθως επιλέγονται τεχνικές cross validation, για την εύρεση του σημείου αυτού.



Σχήμα 2.7. Ποικιλότητα σε σχέση με προκατάληψη

2.7 Δικτυακές επιθέσεις

Κύριος κλάδος, όπου οι ανωμαλίες κάνουν συχνά την παρουσία τους και η εξακρίβωση αυτών αποτελεί ζήτημα κρίσιμης σημασίας και θα μας απασχολήσει στη συνέχεια, αποτελούν τα δίκτυα. Σύμφωνα με πληροφορίες του ATLAS, μιας πρωτοβουλίας επιχειρήσεων για ανώνυμο διαμοιρασμό δεδομένων, το 2017 συνέβησαν παραπάνω από 600,000 DDoS επιθέσεις, με το μέγεθος των δεδομένων που εκτίθενται να παρουσιάζει ανάλογα αύξουσα τάση. Η αύξηση αυτή οφείλεται εν μέρη στην αύξηση των διαθέσιμων εργαλείων που διατίθενται, που διευκολύνουν σε μεγάλο βαθμό την όποια απαιτούμενη προσπάθεια.

Η ταξινόμηση των επιθέσεων αυτών μπορεί να γίνει με πολλούς διαφορετικούς τρόπους και διαφορετικά κριτήρια [16]. Στη συνέχεια θα εξετάσουμε κάποια από τα πιο συχνά είδη επιθέσεων.

2.7.1 DDoS Attack

Οι επιθέσεις αυτές, γνωστές και ως επιθέσεις άρνησης υπηρεσίας, στοχεύουν σε κάποια διαδικτυακή εφαρμογή και επικεντρώνονται στην υπηρεσία που διατίθεται από αυτή. Πραγματοποιώντας υπερβολικά υψηλό αριθμό αιτήσεων για την υπηρεσία

που προσφέρεται, αποπειρώνται να την καταστήσουν μη διαθέσιμη. Το είδος των υπηρεσιών αυτών μπορεί να ποικίλει από τις συναλλαγές σε μία τράπεζα μέχρι μια διαδικτυακή εφαρμογή. Στόχος είναι η παράλυση της δυνατότητας ανταπόκρισης του αμυνόμενου. Το αποτέλεσμα είναι δυσχερέστερο με τη χρήση κατανεμημένων συστημάτων.

2.7.2 SYN Flood

Παρόμοια με την προηγούμενη περίπτωση, κάποιες φορές αναφέρεται ως είδος επίθεσης άρνησης υπηρεσίας, ο επιτιθέμενος στέλνει συνεχώς αιτήσεις SYN με στόχο να καταναλώσει τους πόρους και να κάνει το σύστημα μη αποκρίσιμο. Η επίθεση αυτή βασίζεται στην τριμερή χειραψία του πρωτοκόλλου TCP. Ο επιτιθέμενος μπορεί να μην απαντήσει, ώστε να ολοκληρώσει τη διαδικασία της τριμερούς χειραψίας, ή να προβάλει εσφαλμένη διεύθυνση IP. Οι συνδέσεις αυτές θα μείνουν ανοιχτές στον διακομιστή για κάποια χρονικό διάστημα (συνήθως 20 δευτερόλεπτα), καταναλώνοντας πολύτιμους πόρους.

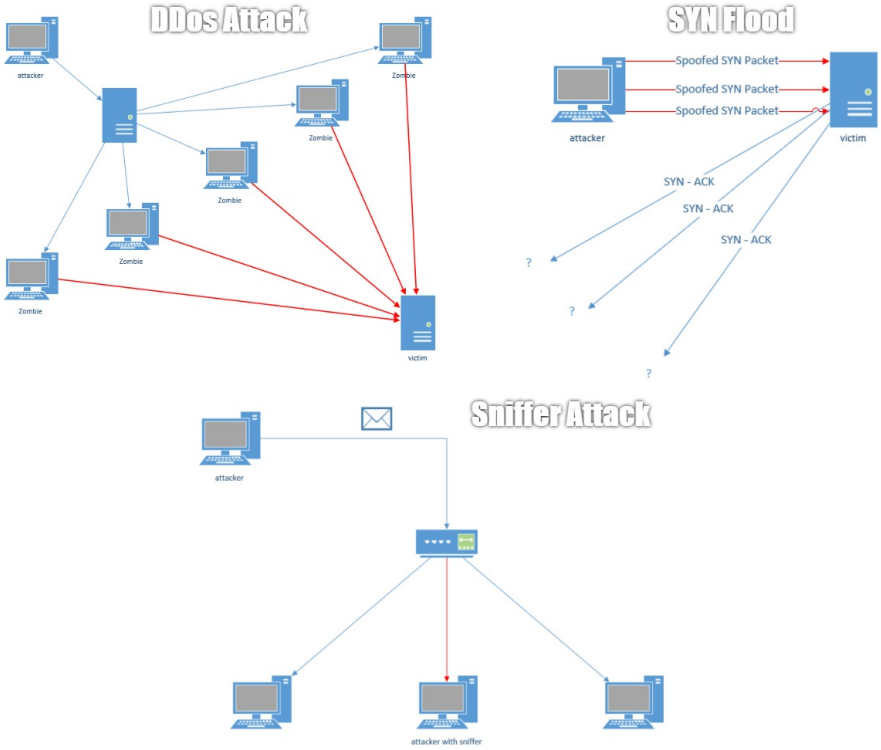
2.7.3 Snifer Attack

Σε περίπτωση που η κίνηση ενός δικτύου δεν είναι κρυπτογραφημένη, εργαλεία που χρησιμοποιούνται για την καταγραφή της κίνησης, μπορούνε να γίνουν σημείο εκμετάλλευσης ενός κακόβουλου χρήστη. Τα περιεχόμενα του πακέτου είναι εκτεθειμένα και μπορούν να γίνουν αντικείμενο ανάλυσης, αποσπώντας κρίσιμες πληροφορίες, όπως κωδικοί.

Άλλα σημαντικά είδη επιθέσεων περιλαμβάνουν Man-In-The-Middle (MITM) attack, IP Address Spoofing Attack, ARP (Address Resolution Protocol) Spoofing Attacks, DNS (Domain Name System) Spoofing Attacks, Phishing and Pharming Spoofing attacks, Password Guessing Attacks, SQL Injection Attacks.

2.8 Intrusion Detection System

Από τα παραπάνω γίνεται εμφανής η ανάγκη ύπαρξης ενός συστήματος ελέγχου της κίνησης ενός δικτύου. Τα συστήματα αυτά συλλέγουν πληροφορίες για την κίνηση του συστήματος και συνήθως καταφεύγουν σε ανάλυση της κίνησης αυτής, ώστε να ορίσουν κάποια βασική συμπεριφορά, καθώς και αποδεκτά όρια και κάποιους κανόνες χρήσης. [32] . Κρίσιμο σημείο αξιολόγησης των συστημάτων αυτών είναι τα ποσοστά false positive και false negative. Τα πρώτα αναφέρονται σε συμβάντα



Σχήμα 2.8. Διαφορετικού τύπου επιθέσεις δικτύων.

που δεν αποτελούν ανωμαλίες αλλά το σύστημά μας εσφαλμένα τα αναγνώρισε ως τέτοια, ενώ τα δεύτερα σε συμβάντα που αποτελούν, αλλά δεν αναγνωρίστηκαν. Εκ πρώτης όψεως, θα θέλαμε όσο το δυνατόν μικρότερο ποσοστό false negative ώστε να αναγνωρίζονται όλες οι ανωμαλίες. Από την άλλη πλευρά ωστόσο μεγάλο ποσοστό false positive, μπορεί να οδηγήσει τον διαχειριστή της πλατφόρμας να δώσει λιγότερη σημασία σε πιθανά κρίσιμα συμβάντα.

Μεγάλη δυσκολία, αποτελεί το συχνά απαγορευτικό μέγεθος των δεδομένων προς επεξεργασία. Οι τεχνικές που θα εφαρμοστούν θα πρέπει να είναι αποτελεσματικές και μικρές σε ανάγκες επεξεργασίας. Το γεγονός αυτό πηγάζει και από ανάγκη επεξεργασίας δεδομένων πραγματικού χρόνου και επομένως άμεση ανταπόκριση σε συμβάντα.

Οι ανάγκες του συστήματος αυτού, απαιτούν ακόμη να είναι ανεκτικό σε σφάλματα, να μπορεί να είναι διαμορφώσιμο αναλόγως με πολιτικές που ορίζονται και να προσαρμόζεται δυναμικά. Επίσης θα πρέπει να έχει τη δυνατότητα κλιμάκωσης σε περισσότερους κόμβους και να παρέχει μια βαθμωτή παροχή της υπηρεσίας του.

Τα συστήματα αυτά μπορούμε να τα κατατάξουμε στις δύο μεγάλες κατηγορίες.

2.8.1 Host-Based Intrusion Detection System

Σε αυτή την κατηγορία ανήκουν ανιχνευτές που λειτουργούν και επενεργούν σε μεμονωμένους κόμβους και συσκευές. Συνήθως είσοδος τους είναι log αρχεία, ακολουθίες κλήσεων συστήματος. Μπορεί να ανιχνεύσει τόσο εξωτερικές όσο και εσωτερικές απειλές, κάτι που δεν είναι δυνατό στην επόμενη κατηγορία. Οι ανωμαλίες εντοπίζονται με βάση μια πολιτική ενεργειών που επιτρέπονται στο χρήστη, καθώς και της ποσότητας των αποπειρών που πραγματοποιήθηκαν. Ορισμένα κριτήρια τα οποία θα πρέπει να λάβει υπόψη το σύστημα αυτό είναι:

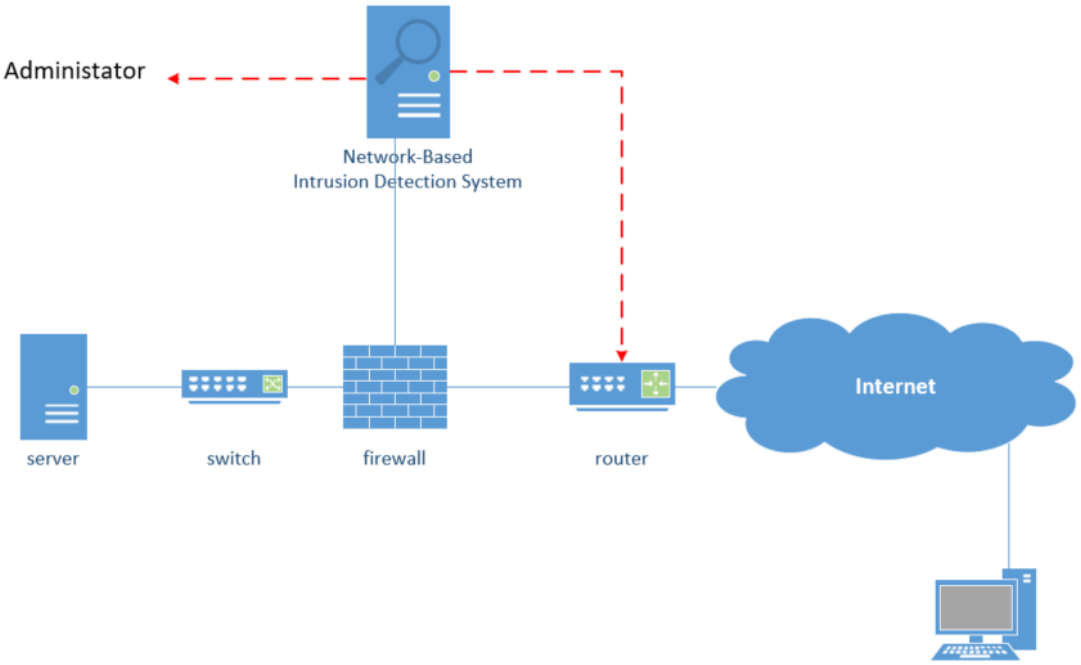
- Ποσοτικοί δείκτες, για παράδειγμα ο αριθμός των συνδέσεων που πραγματοποίησε σε ένα χρονικό διάστημα ή ο αριθμός των ταυτόχρονων συνδέσεων.
- Χρονόμετρα, το χρονικό διάστημα ανάμεσα σε συμβάντα που παρουσιάζουν ενδιαφέρον.
- Στατιστικά μεγέθη, όπως μέση τιμή, τυπική απόκλιση για τη συμπεριφορά του χρήστη.

2.8.2 Network Intrusion Detection System

Αυτά τοποθετούνται σε στρατηγικά σημεία στο δίκτυο, ώστε να παρακολουθούν την κίνηση στα σημεία αυτά. Τα γεγονότα που καταγράφονται, στέλνονται στη συνέχεια σε ένα κεντρικό σημείο επεξεργασίας. Εξετάζει την κίνηση σε πραγματικό ή σχεδόν πραγματικό χρόνο, συνήθως μετά από κατάλληλη δειγματοληψία, ώστε να μην επιβαρύνει σε μεγάλο βαθμό την λειτουργία του υπόλοιπου συστήματος. Αναλόγως με τη λειτουργικότητα του κόμβου μπορούν να εξετάζουν τα δεδομένα σε πρωτόκολλα επιπέδου δικτύου, μεταφοράς και εφαρμογής. Οι αισθητήρες μπορούν να εξετάζουν αυτούσια την κίνηση ή αντίγραφα αυτής, για λόγους γρήγορης απόκρισης των υπόλοιπων λειτουργιών που προσφέρονται.

Σημαντική πρόκληση αποτελεί το γεγονός ότι ο τύπος των ανωμαλιών έχει μεγάλη διάσταση και οι ίδιες οι ανωμαλίες εξελίσσονται γρήγορα με το χρόνο, προσαρμοζόμενες στα νέα συστήματα ανίχνευσης. Ανάλογα με την απόφαση μπορεί να απορρίψει τελείως πακέτα.

Μια τελευταία κατηγορία είναι το κατανεμημένο ή υβριδικό Intrusion Detection System, το οποίο βασίζεται σε ένα συνδυασμό των παραπάνω μεθόδων.

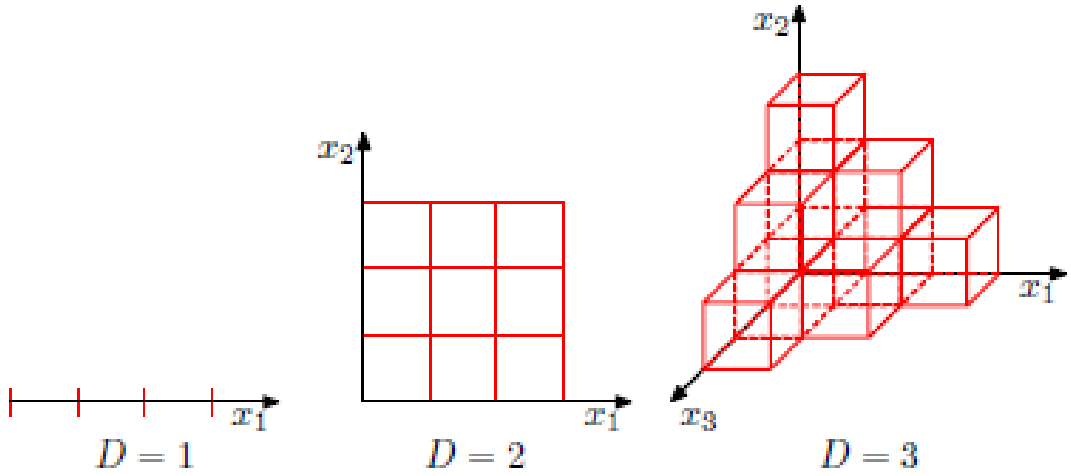


Σχήμα 2.9. Παράδειγμα Network-Based IDS.

2.9 Δεδομένα μεγάλου αριθμού διαστάσεων

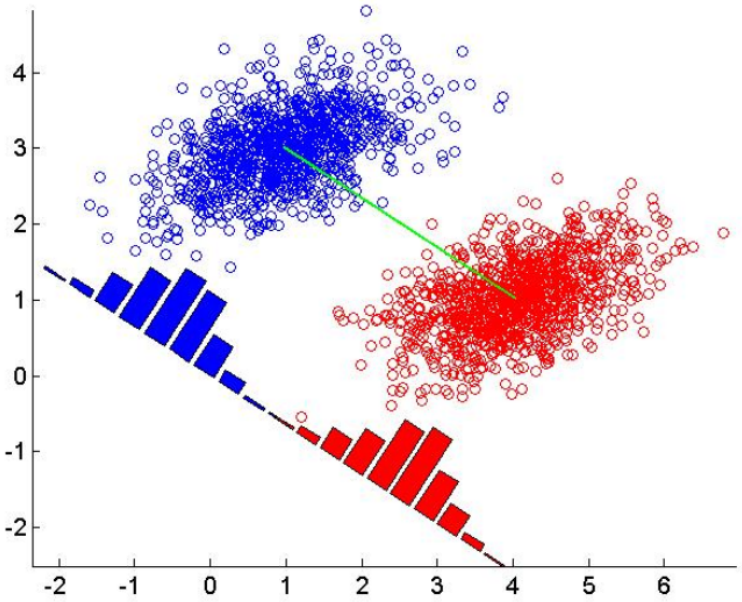
Μία από τις σημαντικότερες προκλήσεις στην εφαρμογή τεχνικών αναγνώρισης προτύπων, αποτελεί η πιθανή μεγάλη διάσταση των δεδομένων. Ήδη από παραδείγματα που αναφέρθηκαν παραπάνω, μπορεί να γίνει προφανής ότι η πιο αποτελεσματική αντιμετώπιση και ως εκ τούτου καλύτερη απόκριση, απαιτεί ως είσοδο μεγαλύτερο αριθμό χαρακτηριστικών, ώστε να επιτρέψει και διαφορετική αντιμετώπιση αυτών.

Κλασσικό παράδειγμα μεθόδων, που αντιμετωπίζουν προβλήματα με την αύξηση της διάστασης των δεδομένων εισόδου, αποτελούν οι μέθοδοι που βασίζονται στη διάσπαση του χώρου χαρακτηριστικών με εφαρμογή ιεραρχικής ομαδοποίησης. Αλγόριθμοι που βασίζονται στη διάσπαση αυτή και παρουσιάζουν μεγάλη δημοτικότητα είναι ο αλγόριθμος Half Space Trees καθώς και ο RS Forest [37, 40]. Κατά τη διάρκεια αυτών, ο χώρος χαρακτηριστικών, όπως φανερώνει και το όνομα τους, διασπάται σε κάθε διάσταση, ανάλογα με τον αριθμό των επαναλήψεων και σε μεγαλύτερο αριθμό περιοχών (buckets). Σε αυτή την περίπτωση, αυξάνοντας το αριθμό των επαναλήψεων, για να διατηρηθεί η ίδια ακρίβεια ανά διάσταση, θα πρέπει να αυξηθεί και εκθετικά ο αριθμός των διαφορετικών περιοχών που θα ορίσουμε στα δεδομένα μας [4].



Σχήμα 2.10. Curse of dimensionality.

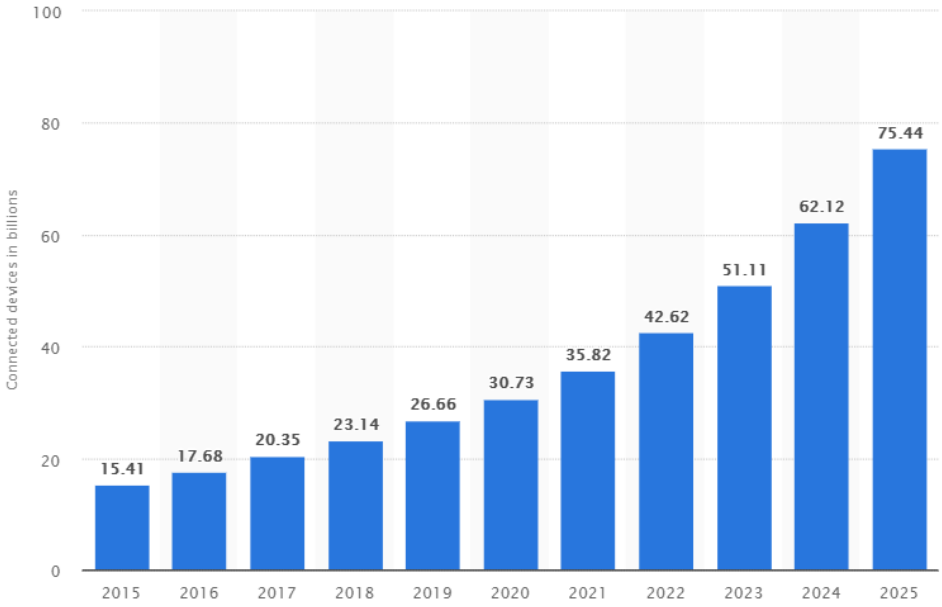
Το φαινόμενο αυτό συνήθως αναφέρεται ως κατάρα πολλών διαστάσεων. Αν και αποτελεί σοβαρό περιορισμό, πολλές τεχνικές έχουν εφαρμοστεί για την καταπολέμηση αυτού. Συνήθως ο χώρος των χαρακτηριστικών μπορεί να μειωθεί αφαιρώντας χαρακτηριστικά που δεν προσθέτουν επιπλέον πληροφορία, ή καταφεύγοντας σε τεχνικές μείωσης των διαστάσεων (PCA, LDA). Απαιτείται ωστόσο προσοχή, καθώς προβολή σε χαμηλότερη διάσταση μπορεί να εισάγει συσχετίσεις μεταξύ ανωμαλιών και μη.



Σχήμα 2.11. Linear Discriminant Analysis

2.10 Ανίχνευση πραγματικού χρόνου

Παρατηρούμε μια τεράστια αύξηση διαθέσιμων δεδομένων στη μορφή ροής δεδομένων πραγματικού χρόνου. Η αύξηση αυτή οδηγείται από την αύξηση των διασυνδεδεμένων συσκευών (IoT). Όπως μπορούμε να παρατηρήσουμε στο Σχήμα 2.12, ο αριθμός των συνδεδεμένων συσκευών στο διαδίκτυο ξεπερνά ήδη τα είκοσι δισεκατομμύρια, ενώ προβλέπεται μια επιταχυνόμενη τάση στα μεγέθη αυτά.



Σχήμα 2.12. Αριθμός διασυνδεδεμένων συσκευών (@Statista 2018)

Τα δεδομένα καταφτάνουν με μια ροή (data stream) και θεωρούμε ότι δεν είναι διαθέσιμα με τυχαία προσπέλαση στο χώρο αποθήκευσής τους, όπως θα συνέβαινε για παράδειγμα με τη χρήση μιας βάσεως δεδομένων. Η χρήση του μοντέλου αυτού, επιφέρει επίσης τις ακόλουθες διαφορές: [2]

- Τα δεδομένα φτάνουν σε πραγματικό χρόνο. Η ροή των δεδομένων αυτών είναι ενεργή με την έννοια ότι δεδομένα καταφθάνουν με εξωτερικές ενέργειες, όπως για παράδειγμα την ύπαρξη κάποιων αισθητήρων και όχι ως αποτέλεσμα ζητήματος που εξέδωσε το σύστημα.
- Όπως προαναφέρθηκε η σειρά με την οποία καταφθάνουν αυτά δεδομένα είναι συγκεκριμένη και δεν είναι υπό τον έλεγχο του συστήματος.
- Δεδομένα υφίστανται επεξεργασία μία φορά, μετά την οποία αυτά αποθηκεύονται ή διαγράφονται. Θεωρούμε ότι το μέγεθος της μνήμης που διαθέτουμε

είναι μικρό σε σχέση με τα δεδομένα αυτά που καταφθάνουν.

- Λόγω υπολογιστικού κόστους και χώρου αποθήκευσης, συνήθως η επεξεργασία που λαμβάνει χώρα είναι προσεγγιστική.

Σημαντική δυνατότητα των εισερχόμενων ροών αυτών είναι να μπορούν να μοντελοποιηθούν ξεχωριστά και να ανιχνεύσουν ανωμαλίες σε πραγματικό χρόνο, αν και η εξαγωγή έμπιστων συμπερασμάτων συχνά αντιμετωπίζει πολλές δυσκολίες. Πολλές φορές δεν αρκεί καν η επεξεργασία δεδομένων σε ομάδες (batches), αλλά απαιτείται ξεχωριστή επεξεργασία κάθε ξεχωριστής τιμής.

Σημαντικός περιορισμός αποτελεί επίσης το κόστος επικοινωνίας. Η ύπαρξη πολλαπλών, κατανεμημένων σημείων ελέγχου της κίνησης, συνεπάγεται πολλαπλά πλεονεκτήματα, καθώς το υπολογιστικό κόστος και ως εκ τούτου και όποια καθυστέρηση εισάγεται μειώνεται. Το γεγονός αυτό είναι πολύ σημαντικό στην περίπτωση ανίχνευσης σε πραγματικό χρόνο. Στον κλάδο αυτό οι περισσότερες τεχνικές χαρακτηρίζονται από μικρό υπολογιστικό κόστος, συχνά καταφεύγοντας σε σημαντική μείωση των διαστάσεων. Παράλληλα ωστόσο η κατανεμημένη αντιμετώπιση αυξάνει το κόστος επικοινωνίας μεταξύ των διαφορετικών τελεστών, που στοχεύει στην ενημέρωση του μοντέλου και στη θεμιτή δυναμική συμπεριφορά.

Έχουν αναπτυχθεί μια σειρά μεθόδων αντιμετώπισης της μεταβαλλόμενης φύσης των δεδομένων, που βασίζονται κυρίως σε προσεγγιστικές μεθόδους, λόγω των δυσκολιών που αναφέρθηκαν προηγουμένως. Οι τεχνικές αυτές μπορούν να κατηγοριοποιηθούν κατά βάση στις παρακάτω κατηγορίες:

- Κινητού παραθύρου (Sliding Window)

  Σε αυτή την περίπτωση δεν εξετάζεται ολόκληρο το ιστορικό για τη λήψη μιας απόφασης, αλλά μόνο ένα παράθυρο. Το παράθυρο αυτό μπορεί να ορίζεται με βάση ένα χρονικό όριο ή με βάση τη συμπλήρωση ενός συγκεκριμένου αριθμού δεδομένων [40, 37] . Με τον τρόπο αυτό επιτυγχάνεται να γίνεται έλεγχος μόνο του πρόσφατου ιστορικού. Κρίσιμη παράμετρος προφανώς είναι η επιλογή του κατάλληλου παραθύρου, καθώς και η επιλογή πλήρης ή μερικής απόρριψης παλαιότερου ιστορικού.

- Επεξεργασία σε τεμάχια (batch processing)

  Αντί για την επεξεργασία ενός δεδομένου με τη σειρά, επιλέγεται η επεξεργασία ενός συνόλου αυτών μαζί. Τα δεδομένα αποθηκεύονται καθώς καταφθάνουν σε προσωρινές δομές πριν μεταβούν στη συνέχεια σε κάποιο στάδιο επεξεργασίας. Η μέθοδος αυτή εμφανίζει μεγάλη δημοτικότητα λόγω της

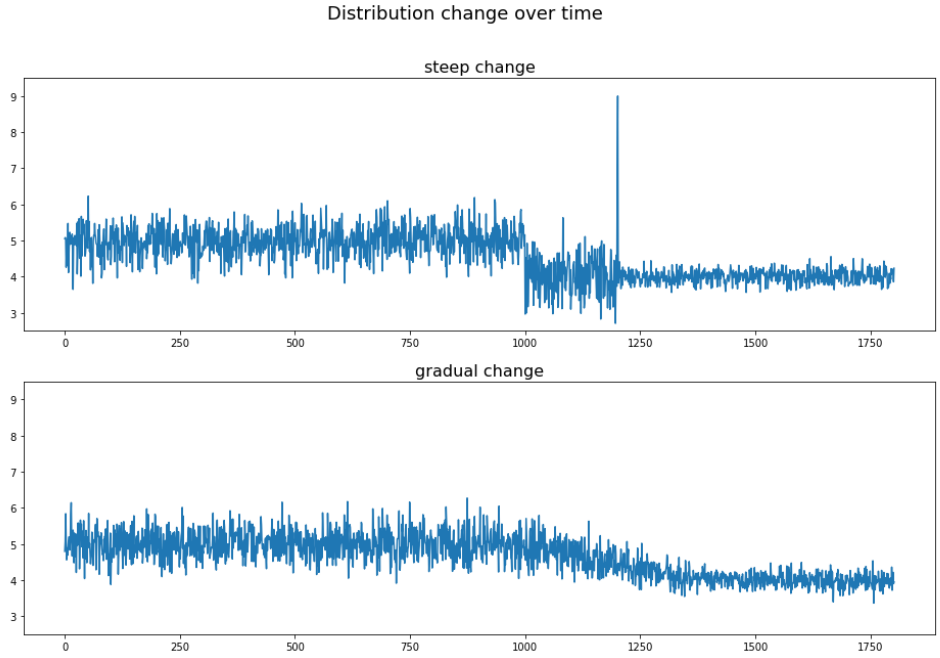
ευκολίας που εμπεριέχει, καθώς και του γεγονότος ότι δε χάνεται κάποια πληροφορία, εις βάρους ωστόσο κάποιου χρονικού διαστήματος για τη συλλογή ενός τεμαχίου. Ανάλογα με το ρυθμό που καταφθάνουν δεδομένα, το ορισμένο μέγεθος ενός τεμαχίου και τον επιθυμητό χρόνο απόκρισης, μια τέτοια καθυστέρηση μπορεί να είναι αποδεκτή ή όχι.

- Δειγματοληψία

  Όταν ο ρυθμός παραγωγής των δεδομένων είναι πολύ μεγάλος σε σχέση με τη διαθέσιμη υπολογιστική ισχύ ή σε περιπτώσεις που μια στοχαστική απόφαση με κάποια ακρίβεια μπορεί να είναι αποδεκτή, χρησιμοποιούνται συχνά ένα υποσύνολο των δεδομένων. Ο ρυθμός της υποδειγματοληψίας αυτής, καθώς και ο τρόπος με τον οποίο συμβαίνει (σε ποιες ροές, για ποιο διάστημα), είναι βασικές παράμετροι.

Σε περίπτωση σημειακών ανωμαλιών (μεμονωμένες περιπτώσεις), το μοντέλο μπορεί να λειτουργεί με τον τρόπο που έχει περιγραφεί μέχρι στιγμής. Πιο προχωρημένες τεχνικές απαιτούνται ωστόσο για περιπτώσεις που παρατηρείται μια παρατεταμένη αλλαγή της κατανομής. Ανάλογα με το πόσο αποκρίσιμο θέλουμε να είναι το μοντέλο μας, μπορούμε να ορίσουμε έναν όρο, με βάση τον οποίο κάθε καινούριο δεδομένο λαμβάνεται υπόψη για την ανανέωση του μοντέλου σύμφωνα με ένα ποσοστό. Η πράξη αυτή ισοδυναμεί με τη δειγματοληψία των νέων δεδομένων που καταφθάνουν. Μεγάλη τιμή του όρου αυτού υποδηλώνει πιθανές γρήγορες αλλαγές κατανομής της εισόδου και επομένως επιθυμία για γρήγορη απόκριση σε αυτές. Αντιθέτως μικρότερος όρος εισάγει μεγαλύτερη αδράνεια στο μοντέλο και δυσκολία αλλαγής της κατάστασής αυτού [41]. Διαφορετικές πρακτικές, υποδεικνύουν η διαφοροποίηση να γίνεται με βάση ένα χρονικό πλαίσιο ή προφανώς συνδυασμό των παραπάνω μεθόδων [14].

Βέβαια, η παραπάνω ανάλυση εισάγει έναν επιπλέον κίνδυνο, αυτό της εσφαλμένης αναγνώρισης αλλαγής στην κατανομής των δεδομένων και με τον τρόπο αυτό πιθανή αποδοχή ανώμαλων συμπεριφορών. Για τη σωστή αντιμετώπιση και τον κατάλληλο συμβιβασμό μεταξύ των καταστάσεων αυτών, θα πρέπει να ληφθεί το κόστος λανθασμένης απόφασης σε κάθε μία περίπτωση σε συνδυασμό με την πιθανότητα εμφάνισης αυτού. Αν ο συνδυασμός αυτός για την περίπτωση εσφαλμένης αναγνώρισης ανωμαλίας (false negative) είναι μεγαλύτερος από τον αντίστοιχο για την περίπτωση αδυναμίας αναγνώρισης σταδιακής μετακίνησης της κατανομής, τότε θα πρέπει να μετακινηθεί αντίστοιχα και το σημείο απόφασης.



Σχήμα 2.13. Αλλαγή μέσης τιμής των δεδομένων (η αλλαγή αυτή μπορεί να γίνει απότομα ή πιο σταδιακά).

Έχει αναπτυχθεί μια ποικιλία διαφορετικών τεχνικών ανίχνευσης ανωμαλιών στον κλάδο. Ένα από τα πιο γνωστά συστήματα, είναια αυτό που έχει αναπτύξει το (Twitter) που βασίζεται στο Seasonal Hybrid ESD [20]

2.11 EM Αλγόριθμος

Βασικός αλγόριθμος που θα χρησιμοποιήσουμε είναι ο αλγόριθμος EM (Expectation - Maximazation). Ο αλγόριθμος αυτός χρησιμοποιείται για τον προσδιορισμό παραμέτρων πίσω από μια κατανομή, ακόμα και αν κάποια από τα χαρακτηριστικά τους είναι άγνωστα. Πρόκειται για έναν επαναληπτικό αλγόριθμο, ο οποίος, όπως δηλώνει και το όνομα του λειτουργεί σε 2 διακριτά βήματα. Στο πρώτο βήμα υπολογίζει την προσδοκώμενη τιμή της ποσότητας log-likelihood, με βάση την τωρινή γνώση για τις παραμέτρους και στο δεύτερο βήμα, υπολογίζει τις παραμέτρους που μεγιστοποιούν την τιμή αυτή που υπολογίστηκε. Στη συνέχεια οι νέες παράμετροι χρησιμοποιούνται εκ νέου επαναληπτικά στο πρώτο βήμα [10, 12].

Υπολογίζουμε την ποσότητα:

$$Q(\theta|\theta^t) = E_{Z|X,\theta^t}[logL(\theta; X, Z)]$$

Θεωρούμε επομένως γνωστές σε αυτό το βήμα τις υποθέσεις μας για την κατανομή

στο προηγούμενο βήμα θ^t.

Στο επόμενο στάδιο βρίσκουμε τις νέες παραμέτρους που μεγιστοποιούν την ποσότητα αυτή:

$$\theta^{(t+1)} = \arg\min_{\theta} Q(\theta|\theta^t)$$

Αξίζει να σημειωθεί ότι έχουν αναπτυχθεί διάφορες παραλλαγές του αλγορίθμου, μερικές από τις οποίες θα χρησιμοποιηθούν στη συνέχεια.

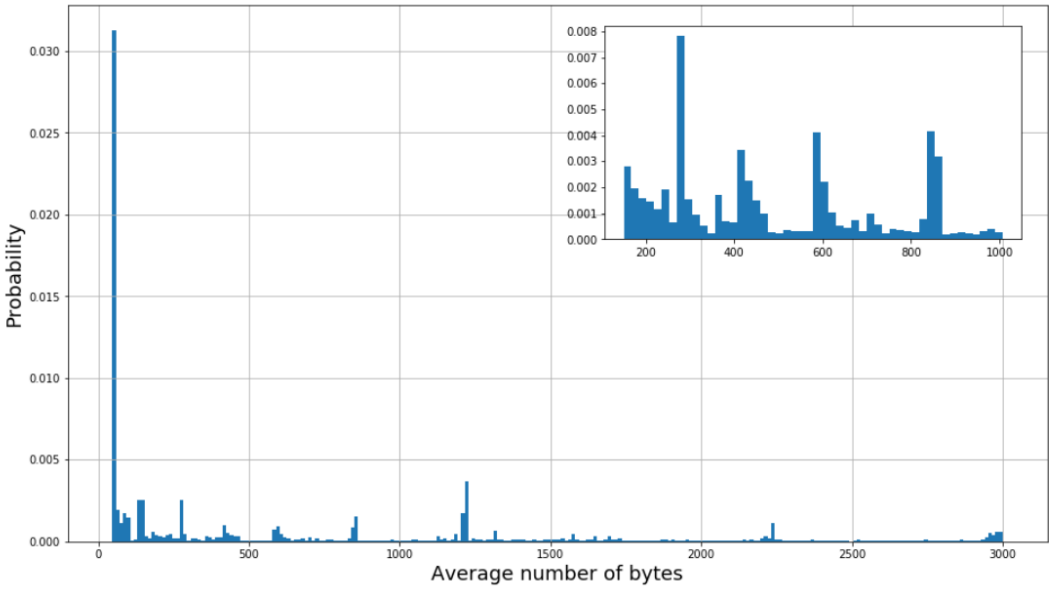
Πριν προχωρήσουμε στη ανάλυση των αλγορίθμων αυτών, θα ήταν δόκιμο να εξετάσουμε για ποιο λόγο θα γίνει χρήση αυτών στην περίπτωσή μας και υπό ποια μορφή.

Ύστερα από παρακολούθηση γύρω από τη μορφή της δικτυακής κίνησης, παρατηρήθηκε ότι αυτή αντιπροσωπεύει σε μεγάλο βαθμό, και επομένως μπορεί να ερμηνευτεί σε ικανοποιητικό βαθμό, κάνοντας χρήση ενός συνόλου από κατανομές (mixtures of distributions) [28].

Το γεγονός αυτό επαληθεύεται σε ικανοποιητικό βαθμό παρατηρώντας ιστόγραμμα δεδομένων κίνησης [22, 21]. Για το λόγο αυτό θα ήταν χρήσιμο η ανάλυση δικτυακής κίνησης από πραγματικά δεδομένα. Το σύνολο των δεδομένων που θα εξετάσουμε και θα μας απασχολήσει, προέρχεται από 58 συνεχόμενες μέρες παρατηρήσεων, από το εσωτερικό δίκτυο των Los Alamos National Laboratory's corporate. Περαιτέρω και πιο λεπτομερής ανάλυση των δεδομένων και των σχέσεων μεταξύ αυτών θα λάβει χώρα σε μετέπειτα στάδιο. Αξίζει να σημειωθεί ότι μέρος των δεδομένων, απαρτίζεται από διαφορετικές ροές που πραγματοποιήθηκαν μεταξύ χρηστών, κατά το χρονικό διάστημα που εξετάζουμε. Στο Σχήμα 2.14 βλέπουμε τον μέσο όρο δεδομένων που ανταλλάσσονται σε κάθε ροή που δημιουργείται. Οι ροές αυτές παρουσιάζουν παρόμοια χαρακτηριστικά με αυτά των κατανομών που μας ενδιαφέρουν (Poisson).

2.11.1 Online EM

Ένα ζήτημα που θα μας απασχολήσει, όπως έχει άλλωστε ήδη αναφερθεί, είναι η δυνατότητα εκμάθησης του αλγορίθμου σε πραγματικό χρόνο, καθώς και η δυνατότητα προσαρμογής αυτού στις μεταβολές που υπόκεινται τα σημεία που εξετάζουμε ή άλλης δυναμικής συμπεριφοράς. Για το λόγο αυτό έχουν αναπτυχθεί τεχνικές που επιτρέπουν την εκπαίδευση του αλγορίθμου αυτού λαμβάνοντας κάθε φορά ένα μέρος των συνολικών δεδομένων (batch). Τέτοιες υλοποιήσεις αναφέρονται στη βιβλιογραφία [26, 25] και θα χρησιμοποιηθούν εκτενώς.



Σχήμα 2.14. Ιστόγραμμα μέσου όρου δεδομένων ανά ροή.

Για την αναπαράσταση των δεδομένων θα γίνει χρήση ενός συνόλου από κατανομές Poisson. Για τον προσδιορισμό μιας νέας παρατήρησης σε ποια από τις παραπάνω κατανομές ταιριάζει προσδιορίζεται ένα ποσοστό συμμετοχής σε κάθε μία από αυτές ως εξής. Αν συμβολίσουμε με $f(.)$ τη συνάρτηση Poisson:

$$f(i|x_t, \theta^k) = \frac{\gamma_i^k f(x_t|i, \theta^k)}{\sum_{l=1}^{m} \gamma_l^k f(x_t|l, \theta^k)}$$

όπου γ είναι οι πιθανότητες εμφάνισεις κάθε κατανομής, δηλαδή a priori γνώση μας κάθε στιγμή για τις κατανομές.

Με βάση αυτές τις συμμετοχές, μπορούμε στη συνέχεια να υπολογίσουμε επαναληπτικά τις νέες παραμέτρους θ^{k+1}. Στην περίπτωση μας αυτές οι παράμετροι αυτές, αναφέρονται στα γ, καθώς και στις παραμέτρους των κατανομών λ.

Ο υπολογισμός αυτός θα γίνει ως εξής:

$$\gamma_i^{k+1} = \frac{1}{n} \sum_{t=1}^{n} f(i|x_t, \theta^k)$$

$$\lambda_i^{k+1} = \frac{\sum_{t=1}^{n} x_t f(i|x_t, \theta^k)}{\sum_{t=1}^{n} f(i|x_t, \theta^k)}$$

Τελικά το μοντέλο θ ανανεώνεται λαμβάνοντας υπόψη τόσο τις προηγούμενες παραμέτρους, όσο και τις νέες υπολογισμένες, με βάση έναν παράγοντα ανανέωσης a_k ως εξης:

$$\theta^{(k+1)} = (1 - a_k)\theta^{(k)} + a_k \Xi(\theta^{(k)}, X_{k+1})$$

όπου η συνάρτηση Ξ, ορίζει την ανανέωση των παραμέτρων όπως ορίστηκε παραπάνω πάνω στα δεδομένα X_{k+1}.

Ο παράγοντας ανανέωσης των παραμέτρων αυτών a_k, εξασφαλίζει σύγκληση στις τελικές παραμέτρους, για συγκεκριμένη κατανομή των δεδομένων αλλάζει, εάν πληροί τις ακόλουθες σχέσεις:

$$\sum_{n=1}^{\infty} a_n = \infty, \sum_{n=1}^{\infty} a_n^2 < \infty$$

Στην πράξη παράγοντες που ικανοποιούν τα παραπάνω είναι $a_k = (k+2)^{-\delta}, \delta \in (0.5, 1]$ ή παράγοντες της μορφής $\frac{1}{k^\delta}$.

2.11.2 Greedy EM

Μια διαφορετική έκδοση του αλγορίθμου είναι η άπληστη (greedy) [39]. Ο αλγόριθμος αυτός, στου οποίου τη σύντομη ανάλυση θα προβούμε σύντομα, μπορεί στην περίπτωσή μας να χρησιμοποιηθεί ως αρχικός προσδιορισμός των παραμέτρων των κατανομών. Το ζήτημα της σωστής αρχικοποίησης είναι σημαντικό και επηρεάζει σε μεγάλο βαθμό το τελικό αποτέλεσμα.

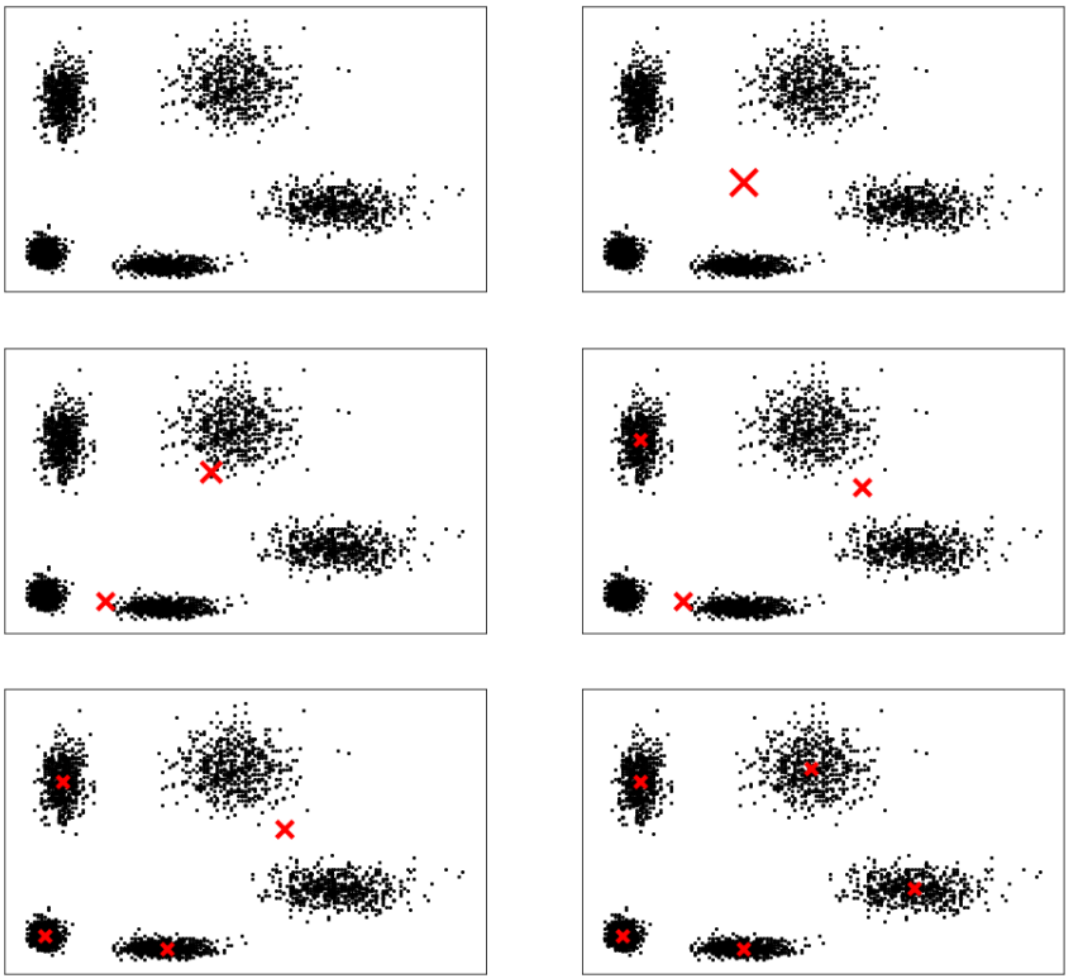
Ο αλγόριθμος αυτός αναλαμβάνει την εκμάθηση των παραμέτρων των κατανομών με ένα άπληστο τρόπο. Ξεκινάει από μία μοναδική αρχική κατανομή, ενώ προσθέτει συνεχώς καινούριες μέχρι έναν επιθυμητό αριθμό k ή έως ότου η προσθήκη καινούριας κατανομής δε βελτιώνει το μέχρι στιγμής μοντέλο. Η ποιότητα επεξήγησης των δεδομένων από το μοντέλο προσδιορίζεται από την ποσότητα $log\_likelihood$.

Ως γνωστόν η συνάρτηση πυκνότητας πιθανότητας στην περίπτωση χρήσης συνδυασμού k κατανομών δίνεται από τον τύπο:

$$f_k(x) = \sum \pi_j \phi(x : \theta_j)$$

όπου με π_j συμβολίζουμε τις a priori πιθανότητες της κάθε κατανομής. Με βάση τα παραπάνω, μπορούμε δημιουργήσουμε μια νέα κατανομή, με στόχο πάντα την αύξηση της ποσότητας $log_likelihood$. Η διαδικασία αυτή μπορεί να φανεί με μεγαλύτερη σαφήνεια στο διάγραμμα 2.15.

Ο προσδιορισμός των παραμέτρων της νέας κατανομής που προστίθεται είναι κρίσιμος για την αποδοτικότητα της μεθόδου. Για λόγους εξοικονόμησης υπολογιστικού κόστους και χρόνου η επιλογή των νέων πιθανών κέντρων μπορεί να γίνει με βάση



Σχήμα 2.15. Επαναληπτική διαδικασία του αλγορίθμου greedy EM.

τα σημεία της κατανομής. Ο προσδιορισμός του βάρους (πιθανότητας) επιλογής κάθε σημείου μπορεί να γίνει με βάση τις τιμές πυκνότητας πιθανότητας που θα προκύψουν για την επιλογή αυτού, σε σχέση με τις παλαιές.

2.12 Bregman divergence

Ο όρος Bregman divergence ή Bregman distance, αποτελεί ποσότητα παρόμοια με απόσταση, με τη διαφορά ότι δεν ικανοποιεί τις ιδιότητες της τριγωνικής ανισότητας και της συμμετρίας που απαιτούνται. Η ποσότητα αυτή ορίζεται ως εξής [5]:

Έστω μια συνάρτηση $\phi\colon S \to \mathbb{R}$, μιας αυστηρώς κυρτή συνάρτηση, παραγωγίσιμη συνάρτηση. Η ποσότητα Bregman divergence που σχετίζεται με τη συνάρτηση ϕ για τα σημεία x και y είναι η διαφορά μεταξύ της τιμής της συνάρτησης ϕ στο σημείο

x και της τιμής του αναπτύγματος Taylor πρώτης τάξης στο σημείο y γύρω από το σημείο x.

$$d_\phi(x, y) = \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle$$

Η πιο απλή και διαδεδομένη τέτοια απόσταση είναι η τετραγωνική Ευκλείδεια απόσταση, η οποία ορίζεται για τη συνάρτηση $\phi(x, y) = \langle x, x\rangle$. Η συνάρτηση αυτή είναι κυρτή και παραγωγίσιμη στο χώρο $\mathbb{R}^d$, οπότε και θα είναι:

$$\begin{aligned} d_\phi(x, y) &= \langle x, x\rangle - \langle y, y\rangle - \langle x - y, \nabla\phi(y)\rangle \\ &= \langle x, x\rangle - \langle y, y\rangle - \langle x - y, 2y\rangle \\ &= \langle x - y, x - y\rangle = ||x - y||^2 \end{aligned}$$

Μία διαφορετική Bregman divergence, την οποία θα χρησιμοποιήσουμε είναι η λεγόμενη Kullback–Leibler divergence. Χρησιμοποιείται για τη σύγκριση δύο κατανομών και περαιτέρω ανάλυση θα λάβει χώρα σύντομα. Αν p και q είναι δύο συναρτήσεις πυκνότητας μάζας πιθανότητας, η αρνητική εντροπία $\phi(p) = \sum_{j=1}^{d} p_j \log p_j$ είναι μια κυρτή συνάρτηση και θα έχουμε:

$$\begin{aligned} d_\phi(p, q) &= \sum_{j=1}^{d} p_j \log p_j - \sum_{j=1}^{d} q_j \log q_j - \langle p - q, \nabla\phi(q)\rangle = \\ &= \sum_{j=1}^{d} p_j \log p_j - \sum_{j=1}^{d} q_j \log q_j - \sum_{j=1}^{d} (p_j - q_j)(logq_j - loge) = \\ &= \sum_{j=1}^{d} p_j \log(\frac{p_j}{q_j}) - \sum_{j=1}^{d} \log e(p_j - q_j) = KL(p||q) \end{aligned}$$

καθώς $\sum_{j=1}^{d} p_j = \sum_{j=1}^{d} q_j = 1$

Μια ενδιαφέρουσα ιδιότητα των Bregman divergence είναι ότι δεδομένου ενός συνόλου από στοιχεία, ο μέσος όρος αυτών ελαχιστοποιεί την απόσταση από το σύνολο των στοιχείων [3]. Η ιδιότητα αυτή δικαιολογεί τη χρήση του μέσου για την αναπαράσταση ενός συνόλου δεδομένων.

2.12.1 Kullback–Leibler divergence

Η ποσότητα αυτή είναι ένα μέτρο του κατά πόσο μια κατανομή πιθανότητας διαφοροποιείται σε σχέση με μία άλλη και συμβολίζεται με $KL(P||Q)$ [24] για δύο

κατανομές P και Q. Ονομάζεται και σχετική εντροπία. Συνήθως χρησιμοποιείται για τη μέτρηση της ποσότητας της πληροφορίας που χάνεται από την επιλογή αντικατάστασης μιας κατανομής πιθανότητας με κάποια άλλη, που ενδεχομένως απαιτεί λιγότερες παραμέτρους για να περιγραφεί [4]. Είναι ένα μέτρο της έκπληξης από την αντικατάσταση αυτή, ή υπό μια διαφορετική ερμηνεία ο αναμενόμενος επιπλέον αριθμός από bits που απαιτείται για την κωδικοποίηση στοιχείων από την κατανομή P μέσω της κατανομής Q.

Σε διακριτό πεδίο ορισμού τυχαίων μεταβλητών, που θα μας απασχολήσουν στη συνέχεια, ορίζεται ως:

$$KL(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

Από την παραπάνω σχέση, αλλά και εννοιολογικά, γίνεται εμφανές ότι για να ορίζεται αν για κάθε i για το οποίο ισχύει ότι $Q(i) = 0$ θα πρέπει να ισχύει επίσης ότι $P(i) = 0$, καθώς ισχύει ότι $\lim_{x\to 0} x \log x = 0$.

Μπορούμε να ορίσουμε επίσης:

$$KL(P||Q) + KL(Q||P)$$

ποσότητα που είναι συμμετρική και μη αρνητική.

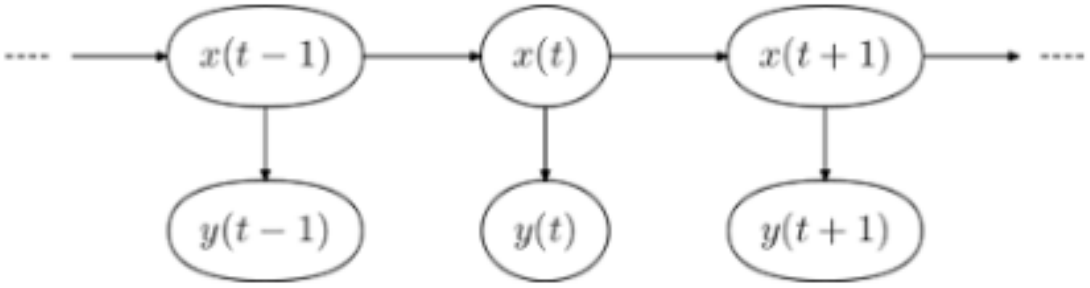
2.13 Κρυφά Μαρκοβιανά μοντέλα

Σε προβλήματα που διακρίνονται από χρονικά μεταβαλλόμενες καταστάσεις, ένας κλασσικός τρόπος αναπαράστασης και εκτίμησης αυτών είναι μέσω των κρυφών Μαρκοβιανών μοντέλων. Η ονομασία τους οφείλεται στο γεγονός ότι δεν μπορούμε να παρατηρήσουμε σαφώς τις καταστάσεις αυτές, αλλά μπορούμε να παρατηρήσουμε κάτι συσχετιζόμενο με την κρυφή κατάσταση αυτή όπως απεικονίζεται στο Σχήμα 2.16.

Βασίζονται, όπως υποδηλώνει και η ονομασία τους, σε Μαρκοβιανές αλυσίδες. Πρόκειται για μία στοχαστική διαδικασία [33], αποτελούμενη από μία ακολουθία καταστάσεων, ενώ η μετάβαση από τη μία κατάσταση σε μία άλλη εξαρτάται μόνο από έναν αριθμό των τελευταίων καταστάσεων αυτών. Για παράδειγμα για ένα τάξης 1 μοντέλο θα ισχύει ότι:

$$P(X_{t+1} = j|X_1 = 1, X_2 = 2, \ldots, X_t = i) = P(X_{t+1} = j|X_t = i)$$

Η χρήση ενός τέτοιου μοντέλου απαιτεί τον προσδιορισμό κάποιων βασικών πα-



Σχήμα 2.16. Κρυφό Μαρκοβιανόμοντέλο. Οι καταστάσεις συμβολίζονται με x, ενώ οι παρατηρησεις που σχετίζονται με αυτές με y.

ραμέτρων που το προσδιορίζουν:

- Πρότερες (prior) πιθανότητες εμφάνισης της κάθε κατάστασης. Αυτές προσδιορίζουν την αρχική κατάσταση, όταν καμία προηγούμενη δεν είναι διαθέσιμη.
- Πίνακας μεταβάσεων, ο οποίος για κάθε κατάσταση προσδιορίζει με πιθανότητα την επόμενη κατάσταση.
- Πίνακας παρατηρήσεων, ο οποίος με βάση την τωρινή κατάσταση, προσδιορίζει ποιες είναι οι πιθανότητες εμφάνισης του κάθε ενδεχόμενου.

Ο τελικός υπολογισμός πιθανότητας σε αυτήν την περίπτωση θα είναι:

$$P(Y, X) = (\prod_{t=1}^{T} P(y_t|x_t)P(x_t|x_{t-1}))P(x_0)$$

Οι τεχνικές εκπαίδευσης και προσδιορισμού πιθανότητας εκβάσεων, στηρίζονται σε τεχνικές δυναμικού προγραμματισμού, όπως οι γνωστοι αλγόριθμοι του Viterbi καθώς και ο backwards-forward.

2.14 Δημιουργία προφίλ

Σκοπός μας είναι η ομαδοποίηση συμπεριφορών των διαφορετικών χρηστών, με στόχο τη δημιουργία διαφορετικών προφίλ χρηστών. Θεωρούμε ότι κάθε χρήστης παρουσιάζει συμπεριφορά που σχετίζεται με τη συμπεριφορά κάποιων εκ των χρηστών που έχουν εμφανιστεί στο παρελθόν. Μπορούμε επομένως, από τα αρχικά δεδομένα, αλλά και από καινούρια που καταφθάνουν να συμπεράνουμε τις διαφορετικές αυτές κλάσεις συμπεριφορών που θεωρούνται και αποδεκτές. Απαιτείται

επομένως η αναγνώριση προτύπων που αναπτύσσονται στα δεδομένα και της μεταξύ τους συσχέτιση. Ανάλογα με το κριτήριο προσδιορισμού, η δημιουργία του κατάλληλου προφίλ μπορεί να αναδειχθεί στις ακόλουθες περιπτώσεις:

- Επιβλεπόμενης ή μη επιβλεπόμενης μάθησης

  Στην πρώτη περίπτωση υποθέτουμε μια συσχέτιση μεταξύ των δεδομένων και ελέγχουμε αυτήν την υπόθεση αν ισχύει. Παράγονται δεδομένα και ελέγχεται η σύνδεση μεταξύ τους. Στη δεύτερη περίπτωση, δεν γνωρίζουμε αρχικά τη σχέση διαφορετικών αντικειμένων, αλλά γίνεται απόπειρα αναγνώρισης πιθανής σχέσης [13].

- Διαφορετικό προφίλ ανά χρήστη ή ανά ομάδα

  Πρόκειται για ένα βασικό πρόβλημα. Από τη μία πλευρά, μπορεί να δημιουργηθεί ξεχωριστό μοντέλο ανά χρήστη με στόχο τον καλύτερο προσδιορισμό της συμπεριφοράς αυτού. Ξεχωριστά μοτίβα που εμφανίζονται στον χρήστη, μπορούν να ενσωματωθούν σε αυτό. Από την άλλη, χρήστες με παρόμοιες συμπεριφορές μπορούν να αντιπροσωπευτούν σε ικανοποιητικό βαθμό από ένα κεντρικό μοντέλο. Ακολουθώντας αυτήν την πρακτική χάνονται πιθανώς ειδικές ανά χρήστη συμπεριφορές, αλλά η μεγαλύτερη πληθώρα δεδομένων επιτρέπει πιθανώς μια πιο ασφαλή εκτίμηση για τις πράξεις αυτού. Επίσης δεν πρέπει να αμελούνται απαιτήσεις μνήμης και υπολογιστικού κόστους. Στην πραγματικότητα, η πιο επιτυχημένη λύση απαιτεί τον κατάλληλο συμβιβασμό μεταξύ των δύο αυτών εναλλακτικών. Πιθανή λύση, απαιτεί την συνύπαρξη των δύο τεχνικών. Μπορεί να δημιουργείται ένα ελαφρύ μοντέλο ανά χρήστη, το οποίο να λαμβάνεται υπόψη στην τελική αξιολόγηση σε συνδυασμό με το προφίλ της ομάδας στην οποία και αυτός ανήκει. Διαφορετικοί ποσοδείκτες μπορούν να προσδιορίζουν το ποσοστό αντιπροσώπευσης του χρήστη από την ομάδα στην οποία τοποθετήθηκε [18].

Σημαντικός παράγοντας αξιοπιστίας της μεθόδου αυτής είναι το κατά πόσο το ομαδικό προφίλ, που έχει δημιουργηθεί, αντιπροσωπεύει αξιόπιστα και τον κάθε επιμέρους χρήστη ή όχι (distributive profile). Για παράδειγμα, ο προσδιορισμός ότι όλα τα άλογα έχουν τέσσερα πόδια, υπό φυσιολογικές συνθήκες, μπορεί να θεωρηθεί αντιπροσωπευτικός για όλο τον πληθυσμό που ανήκει σε αυτήν την κατηγορία.

Ωστόσο η ιδανική αυτή περίπτωση, όπου όλα τα άτομα ενός πληθυσμού μοιράζονται ένα κοινό χαρακτηριστικό, σπάνια εμφανίζεται. Συνήθως η ομαδοποίηση σε μια κατηγορία, εμπεριέχει αρκετές απλοποιήσεις ανάμεσα στο άτομα του πληθυσμού. Για παράδειγμα, μια ενδεχόμενη ομαδοποίηση μπορεί να είναι η κατηγορία των

ψυχοπαθών. Η αναγνώριση ενός ψυχοπαθούς ασθενή, γίνεται με βάση μια σειρά χαρακτηριστικών της προσωπικότητας ενός ανθρώπου. Βαθμολογείται κάθε άτομο σε μια κλίμακα ανάλογα, για παράδειγμα, της κοινωνικής συμπεριφοράς που δείχνει ή την ευαισθησία του σε αδύναμους. Ο προσδιορισμός ενός ανθρώπου ως ψυχοπαθή ή όχι, προκύπτει αν περάσει ένα κατώτερο κατώφλι στον έλεγχο αυτό. Ο μονοσήμαντος αυτός προσδιορισμός (hard clustering), εμπεριέχει προφανώς πολλές απλοποιήσεις.

ΚΕΦΑΛΑΙΟ 3

Μεθοδολογία

3.1 Διατύπωση του προβλήματος

Αρχικά απαιτείται να προβούμε σε μια πιο λεπτομερή περιγραφή του προβλήματος που μας απασχολεί. Θεωρούμε ότι ελέγχουμε την κίνηση που αφορά μια σειρά διαφορετικών χρηστών. Ο έλεγχος αυτός μπορεί να αφορά από μια εταιρεία τηλεπικοινωνιακών συστημάτων μέχρι τον πάροχο μιας δικτυακής εφαρμογής. Κάθε χρήστης παράγει ανά χρονικό διάστημα, μια σειρά μετρήσεων. Πιο συγκεκριμένα, θεωρούμε μια σειρά παρατηρήσεων

$$X_{kij}$$

όπου ο δείκτης k δείχνει το στοιχείο της εκάστοτε παρατήρησης, ο δείκτης i αναφέρεται στη χρονική στιγμή της παρατήρησης και ο δείκτης j στον χρήστη, στον οποίο και οφείλεται η παρατήρηση. Θεωρώντας ότι το σύνολο των διαφορετικών χρηστών που εξετάζουμε είναι N συνολικά:

$$hosts : h_j, j = 1, ..., N$$

Επίσης εξετάζοντας M στο σύνολο διαφορετικές χρονικές στιγμές και θεωρώντας ότι κάθε μία από αυτές αποτελείται από p στο σύνολο χαρακτηριστικά, καταλήγουμε ότι το διάνυσμα χαρακτηριστικών μας X έχει διάσταση $p \times M \times N$. Το X_{*ij} συμβολίζει τις παρατηρήσεις ενός χρήστη μια χρονική στιγμή, ενώ το X_{**j}, αναφέρεται στο σύνολο των παρατηρήσεων ενός αποκλειστικού χρήστη.

Σκοπός μας είναι η ανίχνευση ανωμαλιών στη συμπεριφορά ενός χρήστη. Η ανίχνευση αυτή θα γίνει με δύο βασικούς γνώμονες, το ίδιο το παρελθόν του χρήστη που εξετάζουμε, καθώς και τη συμπεριφορά του ίδιου του χρήστη σε σχέση με τη

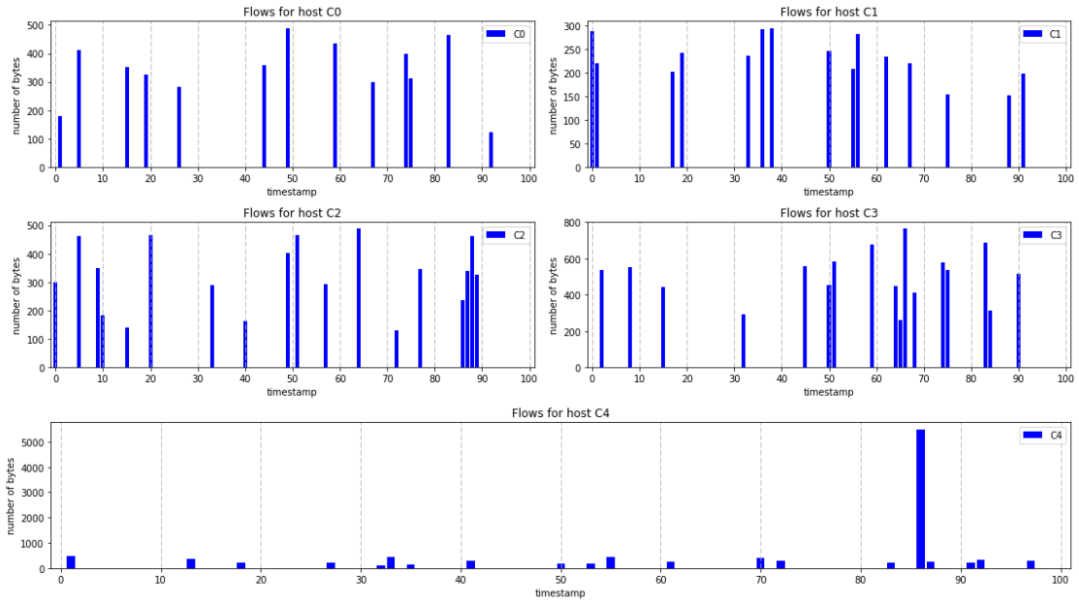
συμπεριφορά των υπολοίπων, ή τουλάχιστον αυτών με παρόμοια συμπεριφορά.

3.2 Κίνητρο

Θεωρούμε ότι έχουμε μια σειρά παρατηρήσεων που αφορούν δικτυακή κίνηση της ακόλουθης μορφής:

$$(timestamp, features)$$

Εξετάζοντας μόνο το μέγεθος των παραπάνω κινήσεων και αγνοώντας προσωρινά τα υπόλοιπα χαρακτηριστικά, μπορούμε να καταλήξουμε στο Σχήμα 3.1:



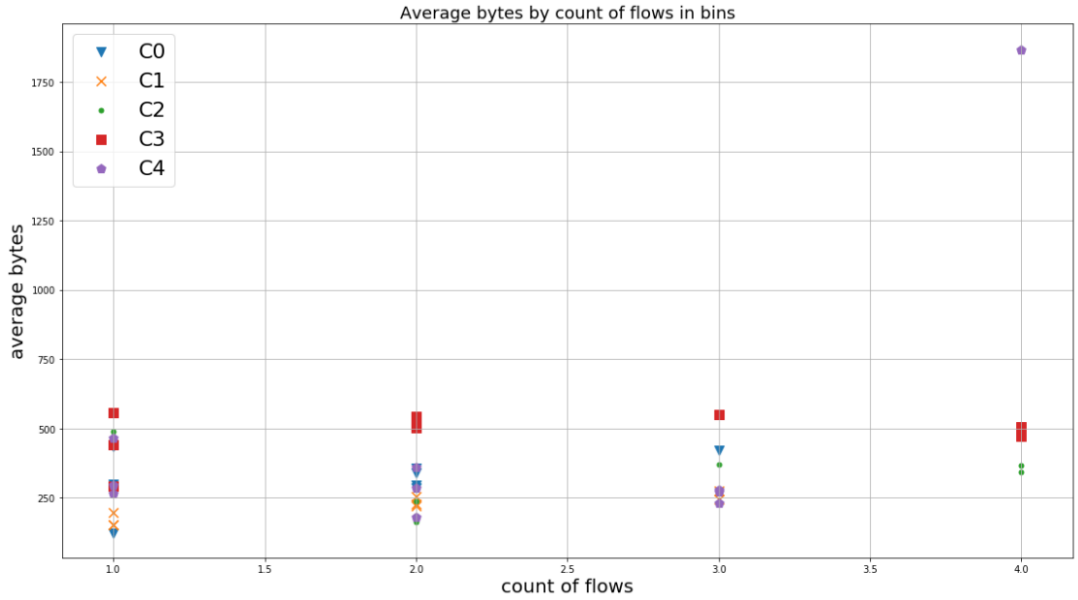
Σχήμα 3.1. Παράδειγμα κίνησης για πέντε διαφορετικούς χρήστες. Η χρονική στιγμή του κάθε συμβάντος είναι στον οριζόντιο άξονα, ενώ στον κάθετο βλέπουμε το μέγεθος της κάθε ροής.

Στο παράδειγμα που εξετάζουμε, όπως αυτό φαίνεται στο διάγραμμα, έχουμε 4 διαφορετικούς χρήστες C0 - C3 οι οποίοι θεωρούμε ότι παρουσιάζουν φυσιολογική συμπεριφορά και ένα επιπλέον χρήστη C4 με παρόμοια χαρακτηριστικά, με εξαίρεση μια χρονική στιγμή, κατά την οποία εμφανίζει μια αναπάντεχη αύξηση της κίνησής του.

Η παραπάνω αναπαράσταση της πληροφορίας δεν είναι ιδανική. Για να καταπολεμήσουμε το γεγονός αυτό, μπορούμε να ομαδοποιήσουμε τα γεγονότα αυτά που συμβαίνουν ανά χρήστη σε χρονικά διαστήματα συγκεκριμένου μήκους. Τότε γίνεται να λάβουμε υπόψη μια σειρά διαφορετικών στατιστικών χαρακτηριστικών, που

αφορούν την κίνηση στο χρονικό αυτό διάστημα που έχουμε λάβει υπόψη.

Στο Σχήμα 3.2 έχουμε ομαδοποιήσει την κίνηση κάθε χρήστη σε χρονικά διαστήματα κρατώντας τον αριθμό των γεγονότων που συνέβησαν ανά χρονική στιγμή, καθώς και το μέσο αριθμό του μεγέθους σε bytes της ροής που αντιστοιχεί στα γεγονότα αυτά.



Σχήμα 3.2. Μέσο μέγεθος και αριθμός ροών ανα χρήστη.

Όπως γίνεται προφανές από το διάγραμμα, οι χρήστες που εξετάζουμε εμφανίζουνε ένα σχετικά χαμηλό αριθμό ροών (0-6) ανά χρονικό διάστημα που εξετάζουμε, με τον μέσο όρο του μεγέθους της ροής να κινείται σε παρόμοια πλαίσια. Η τεχνητή εξαίρεση του κανόνα που έχουμε εισάγει στον τελευταίο χρήστη ξεχωρίζει με μεγάλη ευκολία.

3.3 Τεχικές που βασίζονται στο μέσο όρο

3.3.1 Υπολογισμός

Πρώτη και πιο απλή προσέγγιση είναι ο υπολογισμός μιας μέσης τιμής της κίνησης και ο προσδιορισμός ενός επιτρεπτού ορίου κίνησης γύρω από την τιμή αυτή.

Για ευκολία μπορούμε να υποθέσουμε ότι οι παρατηρήσεις μας αντιπροσωπεύονται από τυχαίες μεταβλητές ανεξάρτητες ακολουθώντας την ίδια κατανομή. Υποθέτοντας ότι ακολουθούν για παράδειγμα Gaussian κατανομή, μπορούμε από την εκτίμηση μέγιστης πιθανοφάνειας (Maximum Likelihood) να υπολογίσουμε αυτήν.

Έχοντας ένα διάνυσμα χαρακτηριστικών μεγεθών p, λαμβάνουμε και αντίστοιχα ένα διάνυσμα μέσων όρων μεγέθους p.

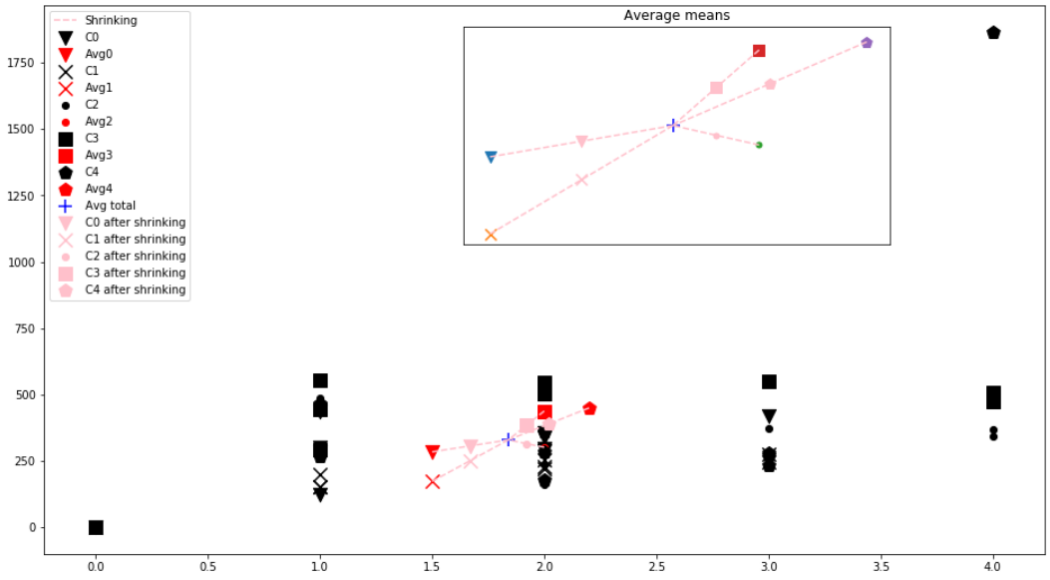
$$\mu = \frac{1}{M}\frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N} X_{*ij}$$

Εκτός από τη μέση τιμή ολόκληρου του συνόλου δεδομένων που διαθέτουμε, μπορούμε να προβούμε στον υπολογισμό της μέσης τιμής ανά χρήστη. Το γεγονός αυτό είναι χρήσιμο για τη σύγκριση της συμπεριφοράς ενός χρήστη σε σχέση με το παρελθόν του. Με τον τρόπο αυτό υπολογίζουμε ακόμα N στο πλήθος διανύσματα μέσων τιμών μεγέθους p το καθένα ως εξής:

$$\mu_j = \frac{1}{M}\sum_{i=1}^{M} X_{*ij}$$

Προφανώς ο ίδιος υπολογισμός μπορεί να λάβει χώρα στον άξονα του χρόνου για τον υπολογισμό μέσης τιμής ανά χρονικό διάστημα.

Ακολουθώντας το παράδειγμα της προηγούμενης ενότητας, στο Σχήμα 3.3 βλέπουμε τις μέσες τιμές κάθε χρήστη ξεχωριστά. Για τα διαστήματα που δεν υπάρχει κίνηση, σημειώνεται το σημείο (0, 0), δηλαδή το σημείο που αντιστοιχεί σε μηδέν ροές για το χρονικό διάστημα αυτό με μηδενικό μέσο όρο μεγέθους μεταδιδόμενων δεδομένων.



Σχήμα 3.3. Μέσος όρος ροής ανά χρήστη και συνολικά

Υπάρχουν δύο διαφορετικοί τρόποι ανίχνευσης ανωμαλιών με βάση τις παραπάνω ποσότητες που υπολογίστηκαν.

- Κανείς θα μπορούσε να ορίσει μια μοναδική καθολικά αποδεκτή συμπεριφορά, θεωρώντας ότι κάθε χρήστης έχει την ίδια ή τουλάχιστον παρόμοια κατανομή. Έτσι η οποιαδήποτε σύγκριση θα γινόταν με βάση το συνολικό μέσο όρο.

- Διαφορετικά, μπορεί κάθε χρήστης να μοντελοποιηθεί ανεξάρτητα και η οποιαδήποτε σύγκριση να συμβαίνει με βάση το δικό του ιστορικό.

Στην πραγματικότητα μια πιο προσεκτική ανάλυση αναδεικνύεται ως η καλύτερη πρακτική, μέσω ενός συνδυασμού των παραπάνω πρακτικών. Μια σωστή ταξινόμηση θα πρέπει να λαμβάνει υπόψη τόσο το συνολικό μέσο όρο του συστήματος όσο και το μέσο όρο του κάθε χρήστη συνολικά. Για το λόγο αυτό εισάγουμε τον όρο αντιστοιχίας shrinkage που δείχνει ακριβώς τη σχέση αυτή, δηλαδή με τι ποσοστό λαμβάνεται υπόψη ο μέσος όρος του χρήστη σε σχέση με αυτόν του συστήματος συνολικά [35]. Η επιλογή της κατάλληλης τιμής αυτής εξαρτάται από πολλούς παράγοντας όπως τον αριθμό των δεδομένων που διαθέτουμε ανά χρήστη και τον κίνδυνο υπερεκπαίδευσης (overfitting). Πρόκειται για το γνωστό δίλημμα ποικιλότητας-προκατάληψης (variance-bias), για το οποίο έγινε αναφορά στο προηγούμενο κεφάλαιο.

Ο μέσος όρος κάθε χρήστη σε αυτήν την περίπτωση προσαρμόζεται ως εξής:

$$\mu'_j = (1 - \eta)\mu_j + \eta\mu$$

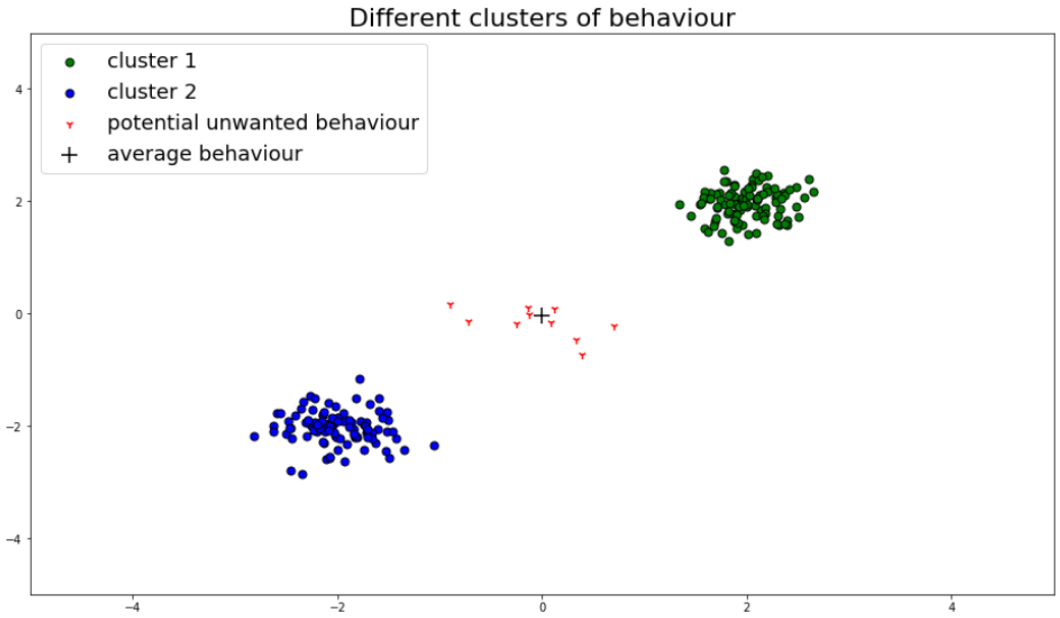
όπου με τον όρο η συμβολίζουμε την ποσότητα shrinkage.

Στο Σχήμα 3.3 παρατηρούμε το μέσο όρο ανά χρήστη συνολικά, καθώς και πως αυτός προσαρμόζεται με τη μεταβολή του όρου shrinkage. Τέλος φαίνονται οι διαμορφωμένοι μέσοι όροι για τιμή αυτού ίση με 0.5.

3.3.2 Προβλήματα που προκύπτουν

Αν και η μέθοδος αυτή διακρίνεται για την απλότητά της και τα γρήγορα αποτελέσματα και συμπεράσματα που μπορούν να παραχθούν από αυτήν, εγκυμονούνται σημαντικοί κίνδυνοι που αποτελούν τροχοπέδη για μια πιο λεπτομερή ανάλυση, όπως φαίνεται στο Σχήμα 3.4.

Η χρήση της προηγούμενης τεχνικής αυτής είναι αποδεκτή, σε περίπτωση μοναδικής κατανομής των δεδομένων. Συνήθως ωστόσο αποδεκτές συμπεριφορές, χαρακτηρίζονται από ένα συνδυασμό κατανομών και μια τάση ομαδοποίησης μεταξύ αυτών [26]. Στο παράδειγμά μας ο μέσος όρος συμπεριφοράς βρίσκεται σε σημείο σχετικά μακριά από τα δεδομένα μας. Το γεγονός αυτό ενισχύεται περισσότερο με την αύξηση των διαστάσεων των δεδομένων.



Σχήμα 3.4. Ομάδες γεγονότων με παρόμοια χαρακτηριστικά.

Για το λόγο αυτό θα υιοθετήσουμε μια διαφορετική τακτική, αυτή της προσέγγισης των δεδομένων με μια σειρά κατανομών.

3.4 Μίξη κατανομών

Για να καταπολεμήσουμε τις αδυναμίες της προηγούμενης μεθόδου θα καταφύγουμε στην αναπαράσταση με ένα σύνολο κατανομών. Λόγω της φύσης των δεδομένων, καθώς και από έρευνες που έχουν πραγματοποιηθεί, καταφεύγουμε στη χρήση ενός συνόλου από κατανομές Poisson. Από τα δεδομένα που διαθέτουμε υπολογίζουμε τις παραμέτρους του μοντέλου με τέτοιο τρόπο ώστε να επιτυγχάνεται εκπαίδευση μέγιστης πιθανοφάνειας (Maximum Likelihood fit) [1]. Δεδομένα που ταιριάζουν σε μεγάλο βαθμό στην μίξη αυτή κατανομών θεωρούνται ως φυσιολογικές περιπτώσεις, ενώ σε αντίθετη περίπτωση πιθανές ανωμαλίες. Το μοντέλο αυτό, σε αντίθεση με άλλες τεχνικές που παρουσιάστηκαν, είναι generative μοντέλο, δηλαδή μπορεί να λειτουργήσει σαν πηγή νέων δεδομένων.

Με πιο αυστηρή διατύπωση θεωρούμε ότι έχουμε ένα μοντέλο M, που αποτελείσαι από ένα σύνολο κατανομών G_i, με συναρτήσεις κατανομής πιθανότητας $f^i(.)$. Η πιθανότητα ένα σημείο X να προέρχεται από την κατανομή αυτή ορίζεται ως:

$$f^{point}(X|M) = \sum_{i=1}^{K} \alpha_i f^i(X)$$

όπου K είναι ο αριθμός των κατανομών από τις οποίες αποτελείται το μοντέλο M.

Θεωρώντας ανεξάρτητα δείγματα μεταξύ τους, η πιθανότητα μια κατανομή D, αποτελούμενη από N το πλήθος μεμονωμένα σημεία, να προέρχεται από το μοντέλο μας ισούται με το γινόμενο των περαιτέρω πιθανοτήτων, δηλαδή:

$$f^{data}(D|M) = \prod_{j=1}^{N} f^{point}(X_j|M)$$

Ο λογάριθμος της παραπάνω ποσότητας L ονομάζεται λογαριθμική πιθανοφάνεια (log likelihood), και ορίζει το πόσο κατάλληλο είναι το μοντέλο που εφαρμόσαμε για την κατανομή. Μια πιο βολική αναπαράσταση της παραπάνω ποσότητας είναι ως εξής:

$$L(D|M) = log(\prod_{j=1}^{N} f^{point}(X_j|M)) = \sum_{j=1}^{N} log(\sum_{i=1}^{k} \alpha_i f^i(X))$$

Σκοπός της εκπαίδευσης που θα ακολουθήσει είναι η ελαχιστοποίηση αυτής της ποσότητας για την καλύτερη προσαρμογή του μοντέλου στα δεδομένα. Μία επιπλέον δυσκολία που εισάγεται με αυτήν την μέθοδο, είναι ότι για κάθε σημείο δε γνωρίζουμε από ποια κατανομή προέρχεται. Για το λόγο αυτό θα πρέπει να αντιστοιχηθεί μια πιθανότητα προέλευσης από την κάθε κατανομή, με βάση τη γνώση του μοντέλου που έχουμε μέχρι στιγμής. Έτσι προβάλλει η ανάγκη ύπαρξης ενός επαναληπτικού αλγορίθμου EM, όπου η μέχρι στιγμής γνώση θα λαμβάνεται υπόψη για τη βελτίωση των παραμέτρων. Αναφορά στα επαναληπτικά βήματα έγινε και στο θεωρητικό υπόβαθρο.

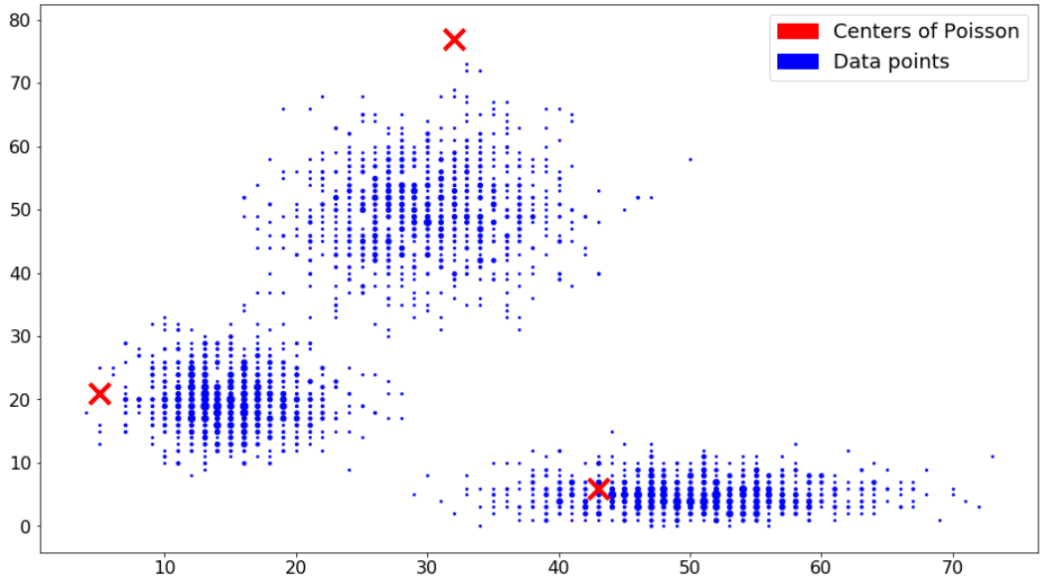
Στο πλαίσιο άφιξης δεδομένων σε πραγματικό χρόνο, έχουν αναπτυχθεί διάφορες τεχνικές υπολογισμού των παραμέτρων αυτών [25, 8]. Βασικός περιορισμός στη χρήση της απλής μορφής του αλγορίθμου EM, αποτελεί η ανάγκη αποθήκευσης των δεδομένων για την επαναληπτική διαδικασία που περιγράφηκε. Για το λόγο αυτό χρησιμοποιούμε μια στοχαστική διαδικασία, όπου οι παράμετροι ανανεώνονται σταδιακά από μέρος των δεδομένων. Η επιλογή της κατάλληλης παραμέτρου ανανέωσης των παραμέτρων σε συνδυασμό με την υπόθεση ότι τα δεδομένα δεν αλλάζουν κατανομή και επομένως ο όποιος θόρυβος εμπεριέχεται θα επαλειφθεί, οδηγεί σε σύγκλιση.

Στο παράδειγμα που ακολουθεί δείχνουμε τη διαδικασία αυτή. Δημιουργούμε σημεία σε δισδιάστατο χώρο, που έχουν προκύψει ως αποτέλεσματα κατανομής Poisson. Τα δεδομένα προέκυψαν από ανεξαρτησία στις διαστάσεις. Πιο συγκεκριμένα τα κέντρα των κατανομών που επιλέχθηκαν ήταν τα (15,20), (30,5) και (50,5). Τονίζουμε ότι η αρχική επιλογή των κέντρων των κατανομών είναι πολύ σημαντική, για να μην συγκλίνουν μεταξύ τους και να αντικατοπτρίζουν το σύνολο των δεδομένων.

Για το σκοπό αυτό επιλέξαμε τον αλγόριθμο k - means++. Τα αρχικά κέντρα διαλέγονται μεταξύ των σημείων των δεδομένων που διαθέτουμε ως εξής:

- Επιλέγεται αρχικά ένα τυχαίο σημείο ως κέντρο μιας κατανομής
- Κάθε νέο κέντρο επιλέγεται τυχαία μεταξύ των υπόλοιπων σημείων υλοποιώντας ένα μηχανισμό ρουλέτας, όπου η πιθανότητα κάθε στοιχείου εξαρτάται από την απόσταση του από το πλησιέστερο κέντρο από αυτά πο έχουν επιλεγεί μέχρι στιγμής.

Ο αλγόριθμος αυτός έχει προσαρμοστεί ώστε αντί της ευκλείδειας απόστασης να χρησιμοποιεί τη συνάρτηση πυκνότητας πιθανότητας Poisson. Στο σχήμα 3.5 βλέπουμε την αρχικοποίηση των κέντρων. Τονίζουμε ότι τόσο στο διάγραμμα αυτό, όσο και σε διαγράμματα που ακολουθούν, το μέγεθος κάθε τελείας που απεικονίζει ένα σημείο δηλώνει και τον αριθμό των σημείων με δεδομένη τιμή. Στο δεδομένο παράδειγμα, έχουμε στο σύνολο 3000 σημεία, 1000 από την κάθε κατανομή. Ο αριθμός αυτός επιλέχθηκε, ώστε να απεικονίζουμε με ένα μικρό αριθμό δεδομένων τη διαδικασία του αλγορίθμου, χωρίς παράλληλα να εισάγεται μεγάλος θόρυβος λόγω της τυχαίας επιλογής των σημείων.

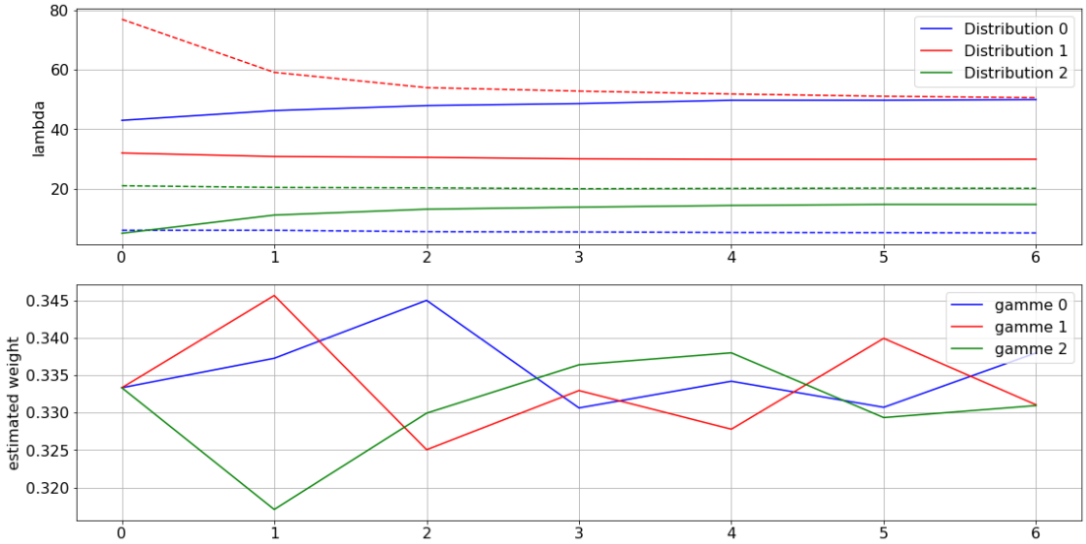


Σχήμα 3.5. Κατανομή σημείων και αρχική τοποθέτηση των κέντρων.

Εκτελούμε τον αλγόριθμο, επεξεργαζόμενοι τα δεδομένα σε ομάδες των 500 σημείων τη φορά, επομένως σε 6 ομάδες (batches). Η συνεισφορά της κάθε ομάδα στις παραμέτρους μειώνεται με τον χρόνο σύμφωνα με την παράμετρο:

$$\frac{1}{n^k}, 0 < k < 1$$

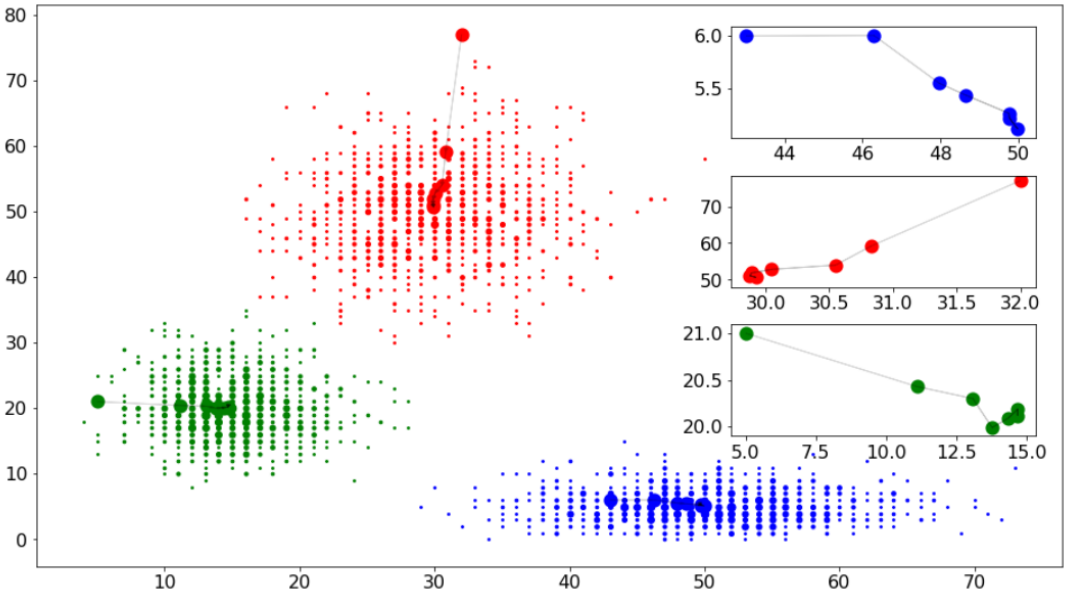
όπου n είναι ο αριθμός της επανάληψης στην οποία βρισκόμαστε. Η επιλογή τιμής της παραμέτρου k θα πρέπει να προσδιοριστεί, στην πράξη η τιμή 0.6 φέρνει σε πολλές περιπτώσεις τα καλύτερα αποτελέσματα [25, 8] από άποψη σύγκλισης. Παρατηρούμε ότι ακόμα και με αυτόν το μικρό αριθμό εκτελέσεων ο αλγόριθμος συνέκλινε ικανοποιητικά στις κατανομές που χρησιμοποιήθηκαν. Αρχικές αποκλίσεις κατά τις πρώτες επαναλήψεις οφείλονται κυρίως στις αρχικοποιήσεις των κέντρων των κατανομών. Λόγω του μικρού αριθμού επαναλήψεων (μόλις 6) παρατηρούμε μια τάση ταλάντωσης γύρω από την επιθυμητή τιμή. Η ταλάντωση αυτή εξαλείφεται με την αύξηση των επαναλήψεων και επομένως τη σταδιακή πτώση του παράγοντα αναβάθμισης.



Σχήμα 3.6. Μεταβολή των παραμέτρων με την επεξεργασία κάθε ομάδας δεδομένων.

Τέλος στο Σχήμα 3.7 εικονίζεται η διαδικασία ανανέωσης των παραμέτρων.

Διαλέξαμε μικρό αριθμό επαναλήψεων, ώστε να φαίνεται με μεγαλύτερη σαφήνεια η διαδικασία ανανέωσης των παραμέτρων αυτών. Επιλογή μικρότερου μεγέθους batch οδηγεί κατά κανόνα σε γρηγορότερη σύγκλιση [25], αλλά δε συμβάλλει στην καθαρή απεικόνιση του παραπάνω παραδείγματος, στη δική μας περίπτωση.



Σχήμα 3.7. Μεταβολή των παραμέτρων σε σχέση με τα σημεία στο χώρο. Το χρώμα των σημείων υποδηλώνει την κατανομή στην οποία ανήκουν.

3.5 Δεδομένα προς επεξεργασία

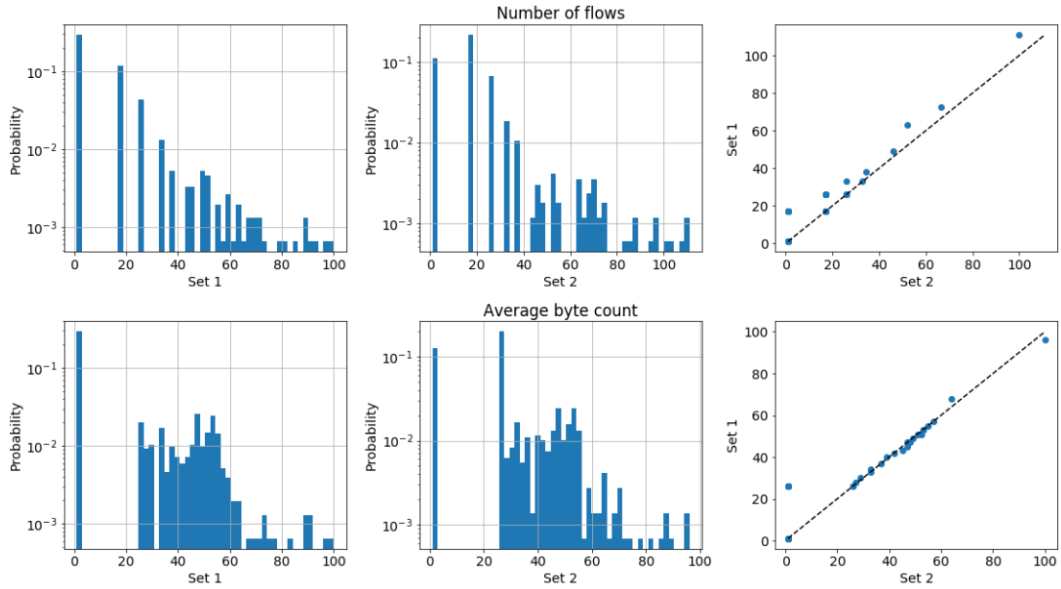
Πριν προχωρήσουμε στην περαιτέρω ανάλυση διαδικασιών και μεθοδολογιών που θα χρησιμοποιηθούν, θα ήταν χρήσιμο να αναφερθούμε στα δεδομένα που διαθέτουμε και την εξάρτηση αυτών από το χρόνο. Απαιτείται η χρήση δεδομένων που αναφέρονται στη διαδικτυακή κίνηση διαφορετικών χρηστών. Όπως προαναφέρθηκε θα γίνει χρήση των ανοικτά διαθέσιμων δεδομένων από το εσωτερικό δίκτυο των Los Alamos National Laboratory's corporate [21, 22].Το σύνολο δεδομένων αυτών, προέρχεται από 58 συνεχόμενες μέρες παρατηρήσεων. Κάθε εγγραφή αναφέρεται σε μια διαφορετική ροή και αποτελείται από τα παρακάτω χαρακτηριστικά:

- χρόνος: αναφέρεται στη χρονική στιγμή έναρξης μιας ροής
- διάρκεια: ο χρόνος που η ροή είναι ενεργή
- πηγή: Ο υπολογιστής-διεπαφή προέλευσης
- προορισμός: Ο υπολογιστής-διεπαφή κατεύθυνσης
- θύρα προορισμού
- πρωτόκολλο
- αριθμός πακέτων
- αριθμός δεδομένων που ανταλλάχτηκαν

Θα ήταν χρήσιμο να ανιχνευτούν εκ των προτέρων οι σχέσεις μεταξύ διαφορετικών χρηστών, καθώς και οι χρονικές συσχετίσεις. Σαν μια πρώτη ανάλυση θα κάνουμε χρήση μόνο του αριθμού των ροών και του μεγέθους των δεδομένων που ανταλλάχτηκαν. Θα μας απασχολήσουν επομένως τα δύο αυτά χαρακτηριστικά. Αρχικά ομαδοποιούμε τις κινήσεις κάθε χρήστη στο χρόνο, ανά ένα σταθερό χρονικό διάστημα. Στη συνέχεια κάθε τέτοιο διαφορετικό χρονικό διάστημα θα το ονομάζουμε εποχή. Η επιλογή της τιμής για το χρονικό διάστημα αυτό θα καλυφθεί σε μετέπειτα ανάλυση. Προς το παρόν, χωρίς βλάβη της γενικότητας, θα επιλέξουμε να ομαδοποιήσουμε την κίνηση ανά ένα λεπτό.

Στο Σχήμα 3.8 βλέπουμε τη σχέση των ιστογραμμάτων διαφορετικών εποχών. Στην περίπτωση αυτή φαίνεται η σχέση μεταξύ διαδοχικών εποχών. Στο Σχήμα 3.9 φαίνεται η ίδια σχέση, δηλαδή η σύγκριση των δεδομένων ανά διαφορετικά διαστήματα, αλλά αυτή τη φορά κάθε σύνολο αποτελείται από ένα σύνολο εποχών, με στόχο

τον προσδιορισμό αλλαγών ανά μεγαλύτερα χρονικά διαστήματα. Από τα διαγράμματα αυτά προκύπτει ότι τα δεδομένα μας έχουν μεγάλη συσχέτιση και σταθερότητα στο χρόνο (τα ίδια πειράματα εκτελέστηκαν και σε μεγαλύτερα χρονικά διαστήματα για να εξακριβωθεί η σχέση αυτή). Το διάγραμμα Q-Q plot [15], δείχνει τη σχέση μεταξύ των σημείων που χωρίζουν τα δεδομένα σε διαφορετικά ποσοστά. Το διάγραμμα αυτό απεικονίζεται σε κάθε περίπτωση ως το δεξιά. Αξίζει να σημειώσουμε ότι κατά μεγαλύτερη πιθανότητα (σχεδόν 70%) κάθε χρήστης εμφανίζει κάθε εποχή μηδενική κίνηση.



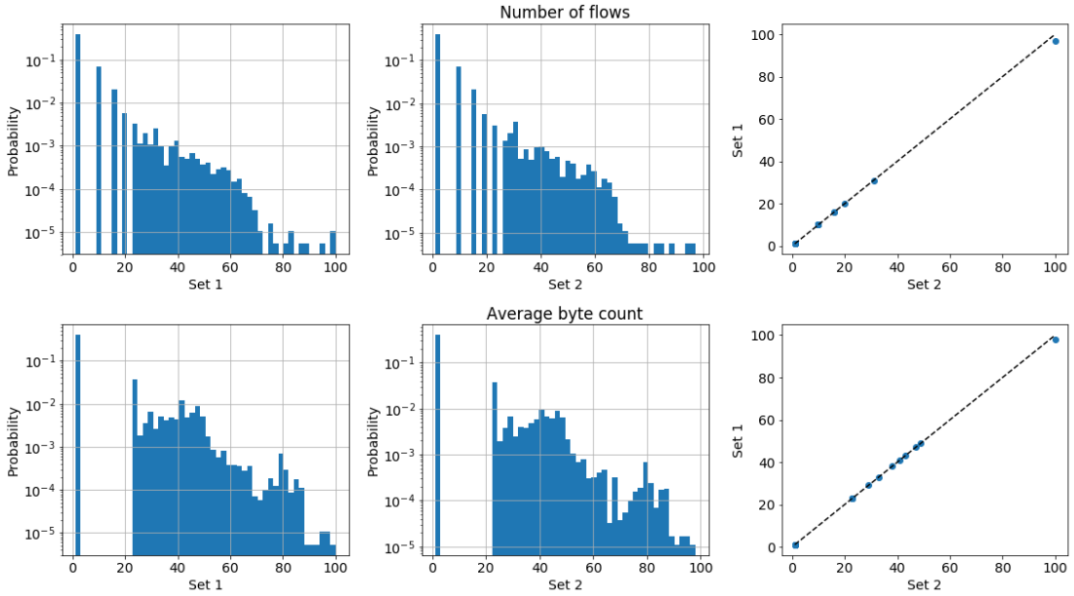
Σχήμα 3.8. Δεδομένα μεταξύ διαφορετικών εποχών, η διάρκεια καθεμίας από τις οποίες είναι ίση με 1 λεπτό. Τα στοιχεία που μας ενδιαφέρουν για κάθε χρήστη είναι ο αριθμός των ροών που αντιστοιχούν σε αυτόν, καθώς και ο αριθμός των δεδομένων ανά ροή.

Μια διαφορετική μέθοδος σύγκρισης δύο διαφορετικών συνόλων δεδομένων είναι το λεγόμενο t-test [23]. Πρόκειται για μια τεχνική στατιστικής ανάλυσης. Κατά τη διάρκεια αυτού συγκρίνονται οι μέσοι όροι δύο κατανομών και βγαίνει ένα συμπέρασμα για το πόσο σημαντικές είναι οι διαφορές μεταξύ των κατανομών αυτών. Με άλλα λόγια, δείχνει αν οι αλλαγές στις κατανομές αυτές μπορεί να είναι τυχαίες.

Υπολογίζεται η ποσότητα t-statistic με την ακόλουθη φόρμουλα:

$$t = \frac{M_x - M_y}{\sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}}$$

όπου M_x M_y είναι οι μέσοι όροι των κατανομών, n_x n_y και S_x S_y δίνονται από τον ακόλουθο τύπο:



Σχήμα 3.9. Δεδομένα μεταξύ διαφορετικών χρονικών διαστημάτων. Κάθε χρονικό διάστημα αντιστοιχεί σε ένα σύνολο εποχών. Στην περίπτωσή μας αντιστοιχεί σε 60 εποχές, δηλαδή συνολικά 1 ώρα καταγεγραμμένης κίνησης.

$$S^2 = \frac{\sum(x - M)^2}{n - 1}$$

Στη στατιστική, ένα μέγεθος που δείχνει τη σημασία των αποτελεσμάτων είναι η ποσότητα p-value. Ανάλογα με την τιμή της ποσότητας αυτής, μπορούμε να βγάλουμε χρήσιμα συμπεράσματα για την υπόθεση που θέλουμε να αποδείξουμε null hypothesis. Το εύρος τιμών κυμαίνεται στο διάστημα από 0 έως 1 και συνήθως επιλέγεται ο παρακάτω εμπειρικός κανόνας:

- Μικρή τιμή (μικρότερη από 0.05) υποδεικνύει ισχυρές ενδείξεις ενάντια της υπόθεσής μας και επομένως πιθανώς απόρριψη αυτής.
- Μεγάλη τιμή (τυπικά μεγαλύτερη από 0.05) υποδεικνύει ασθενείς ενδείξεις απόρριψης της υπόθεσης και επομένως αδυναμία απόρριψης αυτής.
- Τιμές κοντά στην τιμή απόφασης δεν προσφέρουν μεγάλη πληροφορία.

Μέσω της υλοποίησης *ttest_ind* της βιβλιοθήκης *scipy.stats* λαμβάνουμε τα αποτελέσματα όπως φαίνονται στον Πίνακα 3.1 για το ίδιο σύνολο δεδομένων που χρησιμοποιήθηκε και για το Σχήμα 3.9:

Μια διαφορετική τεχνική, το λεγόμενο Kolmogorov–Smirnov test, βασίζεται στη σύγκριση των εμπειρικών συναρτήσεων πιθανοτήτων των κατανομών, οι οποίες υπολογίζονται ως εξής:

	t-statistic	pvalue
number of flows	0.183658694	0.854281412
average number of bytes	0.067537675	0.946153741

Πίνακας 3.1. Αποτελέσματα t-test

	KS-statistic	pvalue
number of flows/average number of bytes	0.002275607	0.992254802

Πίνακας 3.2. Αποτελέσματα Kolmogorov-Smirnov test

Για n ανεξάρτητες παρατηρήσεις X_i:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{[-\inf,x]}(X_i)$$

όπου $I_{[-\inf,x]}(X_i)$ έχει την τιμή 1 αν $X_i \leq x$, διαφορετικά έχει την τιμή 0.

Μέσω του υπολογισμού των εμπειρικών αυτών συναρτήσεων πιθανοτήτων, υπολογίζεται η ακόλουθη ποσότητα:

$$D_n = \sup_x |F_{n1}(x) - F_{n2}(x)|$$

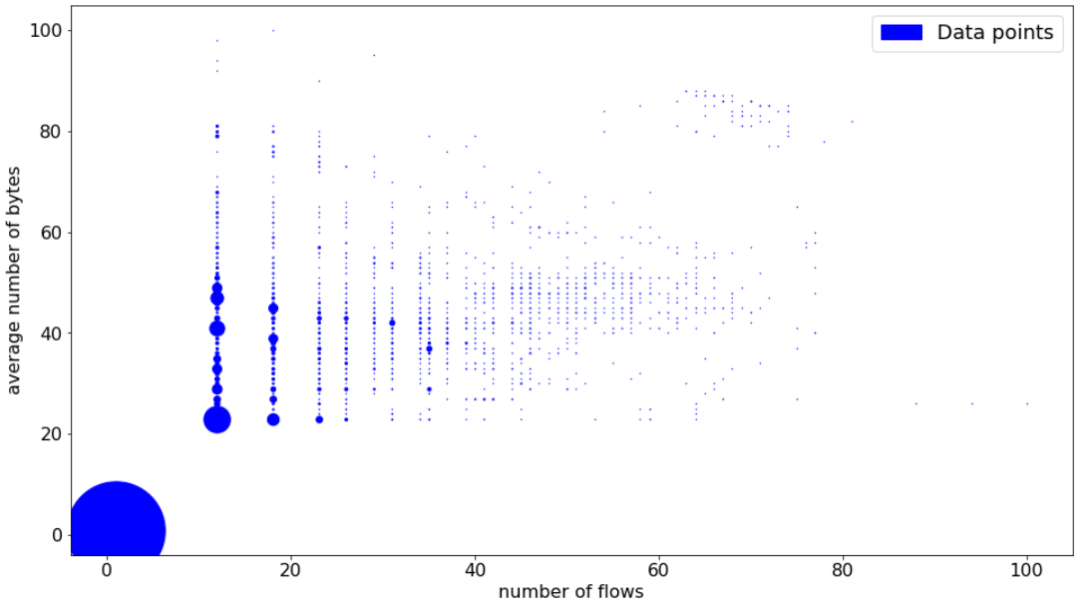
Η μέθοδος αυτή μπορεί να επεκταθεί σε περισσότερες διαστάσεις [29], γεγονός που επιθυμούμε στην περίπτωση μας (επισημαίνουμε ότι όλες οι προηγούμενες μετρήσεις αφορούσαν μεμονωμένες διαστάσεις). Στον Πίνακα 3.2 βλέπουμε τα αποτελέσματα.

Παρατηρώντας τους Πίνακες 3.1, 3.2, καθώς και τα προηγούμενα Σχήματα 3.8, 3.9, μπορούμε να πούμε, ότι τα δεδομένα μας μένουν σταθερά σε ικανοποιητικό βαθμό με το χρόνο.

3.6 Βασικές μέθοδοι

Πριν την περιγραφή του αλγορίθμου που προτείνουμε εξετάζουμε απλούστερες μεθόδους και τα προβλήματα που εμφανίζουν αυτές στην περίπτωσή μας.

Στο Σχήμα 3.10 φαίνεται ένα τμήμα των δεδομένων, που αντιπροσωπεύει κάλλιστα όλα τα δεδομένα. Κάθε κουκίδα αντιπροσωπεύει μία μέτρηση, ενώ το μέγεθος αυτής υποδηλώνει τον αριθμό των μετρήσεων με την ίδια τιμή. Τα δεδομένα έχουν υποστεί επεξεργασία μέσω της συνάρτησης:



Σχήμα 3.10. Στο σχήμα φαίνεται ένα τμήμα των δεδομένων από το Los Alamos National Laboratory. Τα χαρακτηριστικά αυτών είναι ο αριθμός των ροών και ο μέσος αριθμός από bytes που μεταφέρονται. Κάθε σημείο προέκυψε ύστερα από ομαδοποίηση ροών με περίοδο 60 δευτερολέπτων.

$$log(x+1)$$

ενώ στη συνέχεια έχουν υποστεί γραμμική επεξεργασία, ώστε τελικά να βρίσκονται στο διάστημα [0,100].

Μπορούμε να κάνουμε κάποιες σαφείς παρατηρήσεις, οι οποίες και θα μας φανούν χρήσιμες στη συνέχεια.

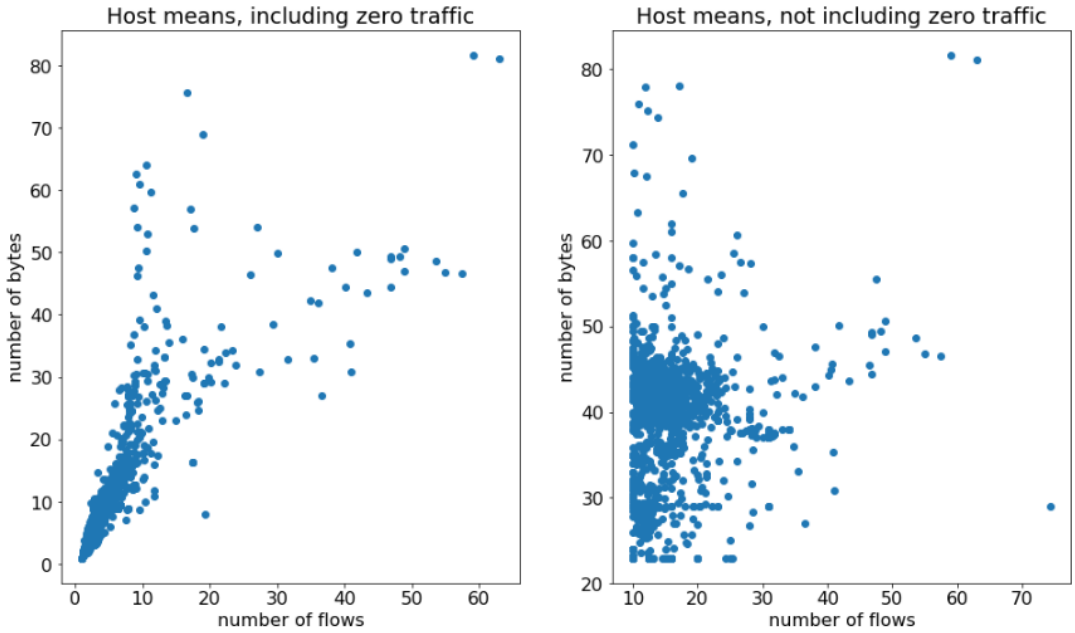
- Το μεγαλύτερο ποσοστό των σημείων βρίσκεται στο σημείο (0, 0), το οποίο και αντιστοιχεί στο σημείο μηδενικής κίνησης στην περίπτωσή μας, δηλαδή για χρήστες που στο διάστημα που εξετάζουμε δεν στέλνουν πακέτα. Συγκεκριμένα πρόκειται για ποσοστό κοντά στο 75
- Το μεγαλύτερο ποσοστό της υπόλοιπης κίνησης εμφανίζεται σε περιοχές με χαμηλό αριθμό ροών και δεδομένων που αποστέλλονται.
- Υπάρχουν μεμονωμένα σημεία μακριά από τα υπόλοιπα δεδομένα. Λαμβάνονται υπόψη και την προεπεξεργασία μέσω της $log(x+1)$, η απόσταση από τα υπόλοιπα σημεία μεταφράζεται σε ακόμα μεγαλύτερη απόκλιση στα αρχικά δεδομένα.

Εξετάζουμε τα απλά μοντέλα που προαναφέρθηκαν στη συνέχεια.

3.6.1 Κατηγοριοποίηση με βάση τον χρήστη

Όπως αναφέρθηκε στο εισαγωγικό απλό παράδειγμα, ένας απλός τρόπος αναπαράστασης των δεδομένων θα ήταν ο υπολογισμός ενός μέσου όρου, στις διαστάσεις που μας ενδιαφέρουν, για τον κάθε χρήστη ξεχωριστά. Στο Σχήμα 3.11 βλέπουμε τα αποτελέσματα κατηγοριοποίησης της κίνησης ανά χρήστη. Όπως προαναφέρθηκε και προηγουμένως κατά μέσο όρο κάθε χρήστης έχει με μεγάλη πιθανότητα (πάνω από 70%) μηδενική κίνηση. Επομένως λαμβάνοντας υπόψη αυτές τις χρονικές στιγμές παρατηρούμε μια μετατόπιση προς το σημείο μηδενικής κίνησης αυτό.

Μια απλή διαδικασία θα αναλωνόταν στη σύγκριση ενός νέου δεδομένου με τον μέσο όρο κίνησης για τον αντίστοιχο χρήστη. Η διαδικασία αυτή διακρίνεται για την απλότητά της.



Σχήμα 3.11. Μέσος όρος κίνησης ανά χρήστη. Στα αριστερά ο μέσος όρος λαμβάνοντας υπόψη στιγμές που ο χρήστης δεν εμφανίζει κίνηση, ενώ στα δεξιά λαμβάνονται υπόψη τέτοιες περιπτώσεις.

Η μέθοδος αυτή ωστόσο μας παρέχει κάποιες χρήσιμες πληροφορίες για τη συνέχεια. Στον Πίνακα 3.3, βλέπουμε το μέσο τετραγωνικό λάθος για τη σύγκριση νέων δεδομένων, με τις μέσες τιμές που υπολογίσαμε για κάθε χρήστη ξεχωριστά. Συγκρίνοντας κάθε νέο δεδομένο με το μέσο όρο του χρήστη στον οποίο και ανήκει αυτό, επιφέρει ένα πολύ μεγάλο τετραγωνικό λάθος. Αντιθέτως, λαμβάνοντας το μέσο όρο κίνησης του χρήστη σε ένα νέο παράθυρο και συγκρίνοντας αυτόν με τον προηγούμενο γνωστό μέσο όρο λαμβάνουμε ένα πολύ μικρότερο. Οι δύο αυ-

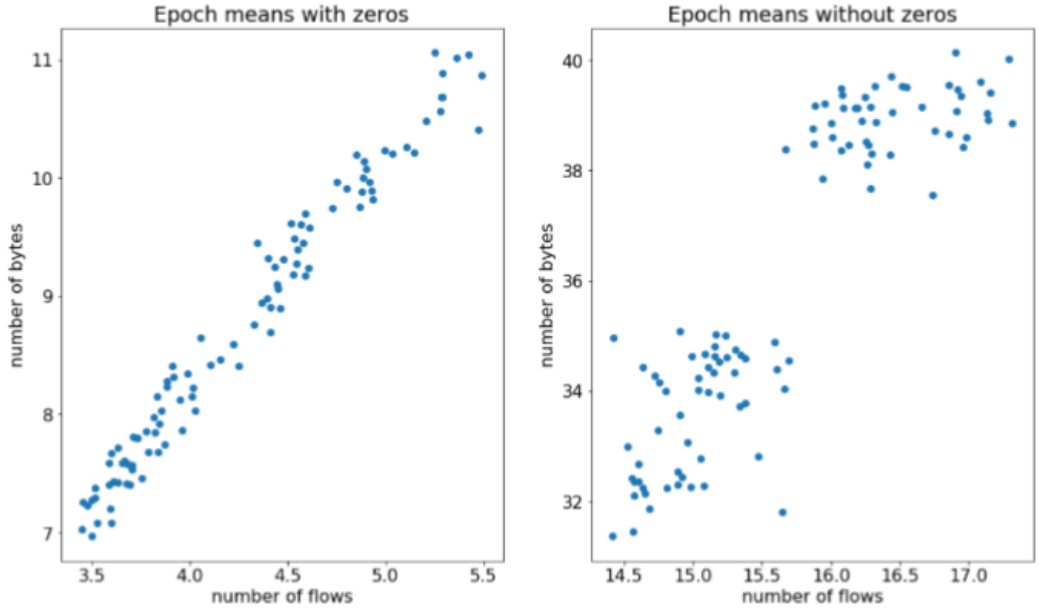
	Μέσο τετραγωνικό λάθος
Σύγκριση κάθε καινούριου δεδομένου με το μέσο όρο κάθε χρήστη	98.10
Σύγκριση του νέου μέσου όρου κάθε χρήστη με τον παλιό	5.26

Πίνακας 3.3. Μέσο τετραγωνικό λάθος για τη σύγκριση με το μέσο όρο κάθε χρήστη στο παρελθόν.

τές παρατηρήσεις επιβεβαιώνουν σε κάποιο βαθμό και τις μέχρι στιγμής υποθέσεις μας. Δηλαδή κάθε χρήστης εμφανίζει μια σχετική σταθερότητα κατά μέσο όρο στο χρόνο, ωστόσο οι παρατηρήσεις μεταξύ τους δεν είναι σε ένα μοναδικό κέντρο συγκεντρωμένες.

3.6.2 Κατηγριοποίηση με βάση την εποχή

Μια διαφορετική προσέγγιση θα ήταν η λήψη του μέσου όρου ως προς το πεδίο του χρόνου. Με τον τρόπο αυτό καταλήγουμε σε μέσους όρους για κάθε μία χρονική στιγμή που εξετάζουμε. Τα αποτελέσματα φαίνονται στο Σχήμα 3.12.



Σχήμα 3.12. Μέσος όρος κίνησης για κάθε εποχή που εξετάζουμε. Αριστερά φαίνεται ο μέσος όρος λαμβάνοντας υπόψη μηδενική κίνηση ανά χρήστη, ενώ στα αριστερά το ίδιο διάγραμμα χωρίς τον υπολογισμό των σημείων αυτών.

Από το διάγραμμα αυτό μπορούμε να βγάλουμε επίσης κάποια σημαντικά συμπεράσματα. Ο λόγος $\frac{number\_of\_flows}{mean(number\_of\_bytes)}$ παραμένει σχεδόν σταθερός, δηλαδή αν και σε

	Μέση τετραγωνική απόσταση
Σύγκριση κάθε δεδομένου με το μέσο όρο κάθε εποχής	201.04

Πίνακας 3.4. Μέσο τετραγωνική απόσταση κάθε δεδομένου από το κέντρο της αντίστοχης εποχής.

κάποιες εποχές παρατηρούμε μεγαλύτερη κίνηση αθροιστικά, ο αριθμός δεδομένων ανά ροή μένει σχεδόν σταθερός. Η σύγκριση κάθε δεδομένου σε σχέση με το μέσο όρο κάθε εποχής δεν μας προσφέρει ικανοποιητική πληροφορία, όπως φαίνεται και από τον Πίνακα 3.4, λόγω της ποικιλότητας που εμφανίζεται ανά εποχή. Διαφορετικά, πιθανή σύγκριση του μέσου όρου μιας εποχής με κάποια προηγούμενη, όπως έγινε σαφές και από το Σχήμα 3.12, παρουσιάζει μικρό σφάλμα, αλλά αδυνατεί να παρουσιάσει με ακρίβεια την πραγματικότητα της κίνησης ανά χρήστη. Η μέθοδος που θα χρησιμοποιήσουμε θα πρέπει να λαμβάνει υπόψη τόσο τις δυσκολίες αυτής όσο και της προηγούμενης μεθόδου.

ΚΕΦΑΛΑΙΟ 4

Βασικές Προσεγγίσεις

Στο σημείο αυτό θα αναλύσουμε κάποιες προσεγγίσεις του προβλήματος και πλεονεκτήματα και μειονεκτήματα της κάθε μίας από αυτές.

4.1 Α.Ομαδοποίηση δεδομένων

Η πιο απλή προσέγγιση που θα μπορούσαμε να χρησιμοποιήσουμε θα ήταν η αναπαράσταση ολόκληρου του συνόλου δεδομένων που διαθέτουμε με μία μοναδική κατανομή. Η προσέγγιση αυτή δεν έχει ιδιαίτερο ενδιαφέρον από μόνη της, αλλά θα αποτελέσει μέτρο σύγκρισης για τις μεθοδολογίες στη συνέχεια.

Το κέντρο της κατανομής αυτής θα είναι απλά ο μέσος όρος των διαθέσιμων δεδομένων. Για τη σύγκριση των διαφορετικών μεθόδων θα αξιοποιήσουμε τη μετρική *log_likelihood*. Η μετρική αυτή παρέχει ένα δείκτη σχετικά με το πόσο αξιόπιστα το μοντέλο που χτίσαμε αναπαριστά τα δεδομένα. τελικώς σκοπός αποτελεί η κατηγοριοποίηση της κίνησης του κάθε χρήστη. Επομένως ο δείκτης αυτός δεν είναι ο πλέον κατάλληλος για το σκοπό αυτό, παρά όλα αυτά παρέχει χρήσιμη πληροφορία για τη σχέση του μοντέλου με τα δεδομένα.

Το σύνολο των δεδομένων μπορεί να αναπαρασταθεί σε μορφή πίνακα, όπως φαίνεται στο Σχήμα 4.2

Σύμβολο	Περιγραφή
K	Αριθμός Κατανομών
N	Αριθμός χρηστών
M	Αριθμός εποχών
J	Αριθμός χαρακτηριστικών

Πίνακας 4.1. Κύριοι συμβολισμοί

Χρήστες / Εποχές	Εποχή 1	Εποχή 2	...	Εποχή M
Χρήστης 1	r_{11}	r_{12}	...	r_{1M}
Χρήστης 2	r_{21}	r_{22}	...	r_{2M}
...	...	...	...	...
Χρήστης N	r_{N1}	r_{N2}	...	r_{NM}

Πίνακας 4.2. Το σύνολο των δεδομένων σε μορφή πίνακα. Συνολικά έχουμε N το πλήθος χρήστες, για κάθε έναν εκ των οποίων διαθέτουμε ένα διάνυσμα χαρακτηριστικών για κάθε μία από τις M εποχές που εξετάζουμε.

Ο υπολογισμός θα γίνει ως εξής:

$$lambda_j = \frac{1}{N \times M} \sum_{i=1}^{N} \sum_{k=1}^{M} r_{ikj}$$

όπου J είναι ο συνολικός αριθμός των χαρακτηριστικών που εξετάζουμε (στην αρχική προσέγγιση μας ο αριθμός των ροών και ο μέσος αριθμός από bytes σε κάθε ροή).

4.2 Β.Ομαδοποίηση δεδομένων και μίξη κατανομών

Στο σημείο αυτό θα εφαρμόσουμε τον αλγόριθμο online - EM (expectation-maximization), όπως αυτός έχει περιγραφεί. Η επιλογή του αλγορίθμου αυτού μπορεί να αιτιολογηθεί ακολούθως.

- Ο online - EM αλγόριθμος δεν απαιτεί αποθήκευση των μετρήσεων.
- Ο όγκος των δεδομένων θα καθιστούσε την επιλογή ενός κλασσικού EM αλγορίθμου υπολογιστικά ασύμφορο. Πιο συγκεκριμένα, ο αριθμός των υπολογισμών θα ανέβαινε ανεπιθύμητα με τον αριθμό των επαναλήψεων που απαιτούνται για σύγκλιση. Στην περίπτωση του online EM, η επαναληπτική διαδικασία στα ίδια δεδομένα αντικαθίσταται από μια επαναληπτική ανανέωση των παραμέτρων του μοντέλου, κοιτώντας ωστόσο κάθε φορά καινούρια δεδομένα, όπως αυτά καταφθάνουν. Εφόσον τα δεδομένα διατηρούν την κατανομή τους, μπορούμε να εγγυηθούμε σύγκλιση όπως αποδείχτηκε στα [26, 25].
- Η ίδια η φύση του προβλήματος επιβάλλει τη χρήση ενός δυναμικά μεταβαλλόμενου μοντέλου. Νέα δεδομένα καταφθάνουν συνεχώς. Ο αλγόριθμος θα πρέπει να είναι σε θέση να αξιοποιεί την καινούρια αυτή πληροφορία, να αναγνωρίζει πιθανές αλλαγές στις κατανομές και να αναπροσαρμόζεται [14].

Η επιλογή του αριθμού των κατανομών που θα κάνουμε χρήση, καθώς και η αρχικοποίηση αυτών αποτελούν βασικές παράμετρους. Για το σκοπό αυτό μπορούμε να χρησιμοποιήσουμε τον αλγόριθμο Greedy EM[39]. Ο αλγόριθμος αυτός ξεκινάει επαναληπτικά από μία μόνο κατανομή και προσθέτει σταδιακά, μέχρι ένα άνω όριο ή μέχρι η ποσότητα *log_likelihood* να μην αυξάνεται κατά ένα προκαθορισμένο ποσοστό. Σε περίπτωση εφαρμογής συγκεκριμένου αριθμού κατανομών, μπορούμε να κάνουμε χρήση της παραλλαγής του αλγορίθμου kmeans ++, για την αρχικοποίηση αυτών.

Τελικά, το *log_likelihood* σε αυτήν την περίπτωση υπολογίζεται ως εξής:

$$ALL = \frac{1}{N \times M} \sum_{i=1}^{N} \sum_{j=1}^{M} log(\sum_{k=1}^{K} \gamma_k f(r_{ij}, l_k))$$

Η μέθοδος αυτή παρέχει τη δυνατότητα να κατηγοριοποιήσουμε κάθε νέο δεδομένο που φτάνει σε μία από τις υπάρχουσες κατανομές. Μας δίνει τη δυνατότητα επίσης για δυναμική αναπροσαρμογή των παραμέτρων της. Η μετέπειτα προσπάθειά μας θα επικεντρωθεί σε μεθόδους που στηρίζονται στον αλγόριθμο αυτό.

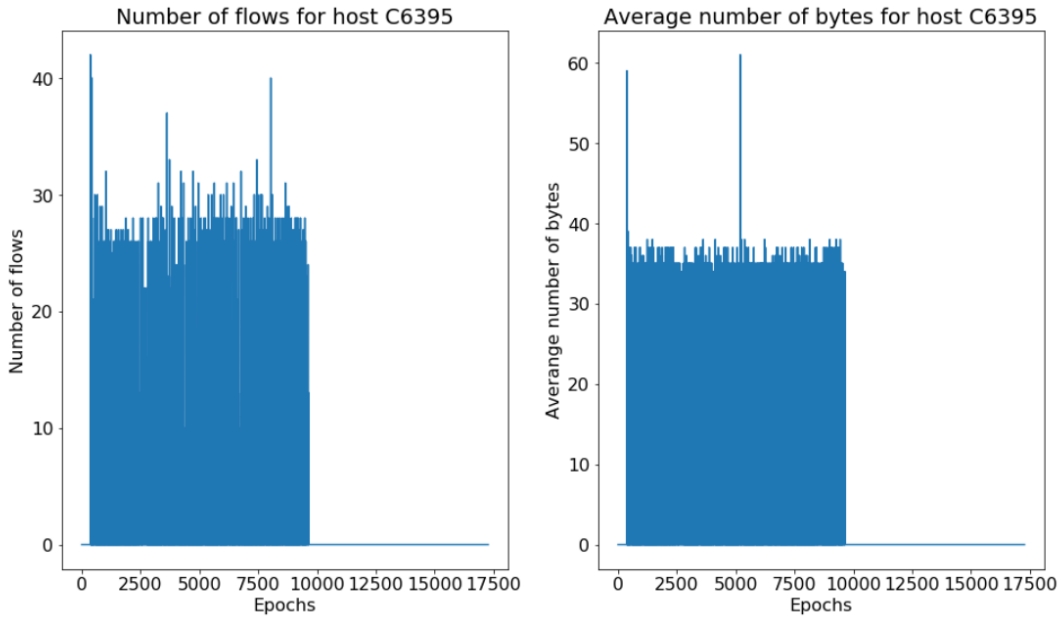
4.3 C.Μίξη κατανομών ανά χρήστη

Μία διαφορετική προσέγγιση προσδιορίζει την ύπαρξη διαφορετικού μοντέλου ανά χρήστη. Δημιουργούμε μια μίξη κατανομών ανά χρήστη, η οποία και αναμένουμε να προσδιορίζει με μεγαλύτερη ακρίβεια τη συμπεριφορά του.

Στην περίπτωση αυτή υπάρχει μεγαλύτερη ανάγκη για σωστή αρχικοποίηση των παραμέτρων. Μεγαλύτερος είναι επίσης ο κίνδυνος υπερεκπαίδευσης (overfitting). Πιο συγκεκριμένα ο αριθμός των δεδομένων ανά χρήστη είναι σημαντικά μικρότερος. Επίσης η συμπεριφορά ενός χρήστη μπορεί να μεταβάλλεται έντονα μεταξύ συγκεκριμένων καταστάσεων, γεγονός που δεν επιτρέπει τη σωστή εκπαίδευση του αλγορίθμου.

Ας θεωρήσουμε για παράδειγμα την ακόλουθη συμπεριφορά ενός χρήστη η οποία φαίνεται στο Σχήμα 4.1. Ο χρήστης αυτός εμφανίζει κάποια κίνηση για έναν αριθμό εποχών, ενώ στη συνέχεια σταματά να εμφανίζει την όποια κίνηση. Λόγω της φύσης του προβλήματος και του αλγορίθμου που χρησιμοποιούμε, η δυναμική ανανέωση των βαρών θα τείνει να αναπροσαρμόσει τις παραμέτρους των κατανομών, ώστε να ανταποκρίνονται στη νέα αυτή συμπεριφορά του χρήστη. Με τη μεταβολή αυτή, επανεμφάνιση προηγούμενης συμπεριφοράς στο νέο ανανεωμένο μοντέλο θα φαντάζει ως διαφορετική από την αναμενόμενη.

Το φαινόμενο αυτό εικάζουμε ότι καταπολεμείται στην περίπτωση χρήσης μοναδικής μίξης κατανομών για όλους τους χρήστες, καθώς κατά μέσο όρο μεταξύ διαφορετικών ομάδων δεδομένων που εξετάζουμε (batches) θεωρούμε ότι διατηρούνται οι κατανομές (όπως δείξαμε πειραματικά στο προηγούμενο κεφάλαιο).



Σχήμα 4.1. Παράδειγμα κίνησης ενός χρήστη, ο οποιός μετά από κάποιο χρονικό διάστημα σταματά να είναι ενεργός.

Τέλος η μεθόδος αυτή καθιστά δύσκολη τη σύγκριση της κίνησης μεταξύ διαφορετικών χρηστών. Παρουσιάζει ωστόσο ενδιαφέρον για τη σύγκριση μελλοντικών τεχνικών.

4.4 D.Κρυφό Μαρκοβιανό μοντέλο με ομαδοποίηση δεδομένων

Σημαντική παράμετρος για την κατηγοριοποίηση των χρηστών, είναι όχι μόνο η κατανομή στην οποία ανήκει η κίνησή τους, αλλά και μεταβάσεις μεταξύ των διαφορετικών κατανομών αυτών. Σκοπός μας στη συνέχεια είναι να αναπαραστήσουμε τη δυνατότητα μετάβασης μεταξύ των διαφορετικών κατανομών για κάθε χρήστη. Αν για παράδειγμα ένας χρήστης εμφανίζει συχνά μια αλλαγή από μία κατάσταση i σε μια άλλη κατάσταση j, θα θέλαμε να επιτρέψουμε παρόμοιες αλλαγές στο μέλλον. Τη διαδικασία αυτή μπορούμε να αποδώσουμε με τη χρήση ενός Μαρκοβιανού μοντέλου.

Παρομοίως με την περίπτωση 4.2, αναπαριστούμε τα δεδομένα με τη χρήση του ίδιου συνόλου από κατανομές. Κατά τη διαδικασία της εκπαίδευσης οι μεταβάσεις μεταξύ των διαφορετικών κατανομών, αναπαρίστανται με τη χρήση ενός πίνακα μεγέθους $K \times K$ (θυμίζουμε ότι K είναι ο αριθμός των κατανομών). Το στοιχείο ij του πίνακα αυτού δείχνει τη πιθανότητα ότι με τωρινή κατάσταση i, η επόμενη κατάσταση θα είναι η j. Αξίζει να σημειώσουμε ότι οι πιθανότητες αυτές είναι εμπειρικές και έχουν προέλθει από μία διαδικασία σκληρής ομαδοποίησης (hard clustering). Κάθε σημείο αντιστοιχίζεται στην κατανομή στην οποία ανήκει με τη μεγαλύτερη πιθανότητα, ενώ οι τιμές του πίνακα μετάβασης που δημιουργείται είναι απλά ο μέσος όρος των σημείων αυτών ως εξής:

$$Transition\_matrix_{ij} = \frac{transitions\_from\_state\_i\_to\_state\_j}{\sum_{k=1}^{K} transitions\_from\_state\_i\_to\_state\_k}$$

Βασιζόμαστε στην ισχυρή σύνδεση μεταξύ του expectation maximization αλγορίθμου και του hard k-means. Τεχνικές που αναλώνονται σε Hidden Maarkov Models με περιορισμούς απαιτούν μεγάλο υπολογιστικό κόστος, που δεν επιτρέπει το πρόβλημά μας, ενώ είναι πέρα από το πλαίσιο αυτής της δουλειάς.

Μέσω αυτής της διαδικασίας παράγεται ένα Μαρκοβιανό μοντέλο. Κάθε τέτοιο χαρακτηρίζεται από κάποιες παραμέτρους:

- Ένας πίνακας π μεγέθους K που υποδηλώνει την πρότερη γνώση μας για κάθε κατανομή.
- Έναν πίνακα a_{ij} μεγέθους $K \times K$ που υποδηλώνει τη πιθανότητα μετάβασης από κάθε κατάσταση σε μία άλλη.
- Έναν πίνακα b_{il} μεγέθους $K \times L$, όπου L είναι οι πιθανές παρατηρήσεις. Κάθε τιμή του πίνακα αυτού δείχνει τη πιθανότητα κάθε παρατήρησης με δεδομένο την κατάσταση στην οποία βρισκόμαστε.

Στην περίπτωσή μας κάνουμε χρήση ενός Μαρκοβιανού μοντέλου γιατί έχουμε ουσιαστικά κλέψει παριστάνοντας ότι τα αποτελέσματα του clustering αλγορίθμου είναι παρατηρήσιμες καταστάσεις. Αλλά δεν είναι απευθείας παρατηρήσιμες, αλλά αναφέρονται μέσω του online EM αλγορίθμου που παρουσιάσαμε. Οπότε αν και αλγοριθμικά υλοποιούμε Μαρκοβιανό μοντέλο, θεωρητικά η διαδικασία του online EM αλγορίθμου σε συνδυασμό με το Μακροβιανό μοντέλο αποτελούν προσέγγιση κρυφού Μαρκοβιανού μοντέλου.

Για μια στοιχειώδη σύγκριση μεταξύ των μεθόδων θα πρέπει να υπολογίσουμε την ποσότητα *log_likelihood*.

Μια απλή λύση στο παραπάνω πρόβλημα είναι η ακόλουθη:

1. Επεξεργαζόμαστε ένα δεδομένο κάθε χρονική στιγμή.

2. Το δεδομένο αυτό ανήκει σε κάποιον χρήστη. Από το ιστορικό του κάθε χρήστη μπορούμε να βρούμε ποια ήταν η τελευταία κατάσταση (κατανομή) στην οποία και αυτός βρισκόταν.

3. Βρίσκουμε σε ποια κατάσταση βρίσκεται το νέο δεδομένο με μεγαλύτερη πιθανότητα υπολογίζοντας τις ποσότητες $\gamma_i \times f(x, l_i)$ για κάθε μία από τις K κατανομές.

4. Με βάσεις αυτές τις ποσότητες μπορούμε να βρούμε σε ποια κατάσταση θα βρίσκεται με μεγαλύτερη πιθανότητα $\underset{k}{\text{argmax}}\gamma_i \times f(x, l_i)$

5. Υπολογισμός του νέου *log_likelihood*

$$log(\sum_{k=1}^{K}(transition\_matrix[previous\_state][k] * f(x, l_k)))$$

   όπου $transition\_matrix[previous\_state]$ υποδηλώνει την πιθανότητα κάθε επόμενης κατάστασης με βάση την τωρινή.

Είναι σημαντικό να τονίσουμε ότι η κατανομή στην οποία κατηγοριοποιείται το κάθε δεδομένο υπολογίζεται μόνο με βάση την αρχική μίξη Poisson που χρησιμοποιήθηκε. Στη συνέχεια, μόνο για τον υπολογισμό του *log_likelihood* τα κέντρα των κατανομών μένουν τα ίδια, αλλά οι πιθανότητες κάθε κατανομής αλλάζουν με βάση την προηγούμενη κατάσταση στην οποία βρισκόμαστε.

4.5 Ε.Κρυφό Μαρκοβιανό μοντέλο ανά χρήστη με ομαδοποίηση δεδομένων

Στην περίπτωση αυτή κάνουμε χρήση της ίδιας ιδέας με την προηγούμενη μεθοδολογία,με την εξαίρεση ότι ο πίνακας μεταβάσεων χαρακτηρίζει κάθε χρήστη ξεχωριστά. Η περίπτωση αυτή παρουσιάζει ομοιότητες με την περίπτωση 4.3, όσον αφορά την προσαρμοσμένη σε κάθε χρήστη αντιμετώπιση. Η μέθοδος μας επιτρέπει τη σύγκριση των χρηστών μεταξύ τους όπως θα διαπιστώσουμε στη συνέχεια.

4.6 F.Ομάδες κρυφών Μαρκοβιανών μοντέλων με ομαδοποίηση δεδομένων

Στην προσπάθεια κατηγοριοποίησης των μεταβάσεων αυτών, βασιζόμαστε στην προηγούμενη μέθοδο και θα προσπαθήσουμε να δημιουργήσουμε ομάδες (clusters) αυτών των πινάκων μεταβάσεων. Για το σκοπό αυτό απαιτείται ο προσδιορισμός δύο μεγεθών:

- Η απόσταση που θα χρησιμοποιηθεί για τη σύγκριση των διαφορετικών πινάκων.
- Ένας τρόπος εύρεσης του μέσου-κέντρου που θα αντιπροσωπεύει την κάθε ομάδα.

Θα χρησιμοποιήσουμε την απόσταση Kullback - Leibner. Στη βιβλιογραφία, έχουν αναπτυχθεί τεχνικές ομαδοποίησης κρυφών Μαρκοβιανών μοντέλων, με την απόσταση αυτή, η πλειονότητα των οποίων βασίζεται σε προσεγγιστικές τεχνικές μέσω μεθόδων Monte Carlo. Το αρνητικό το τεχνικών αυτών, είναι ότι απαιτούν μεγάλο αριθμό επαναλήψεων, ώστε να έχουμε πεποίθηση για τη σύγκλιση και το τελικό αποτέλεσμα.

Μια διαφορετική τεχνική, που αναλώνεται στον υπολογισμό ενός άνω φράγματος της απόστασης που ζητείται σε πολύ μικρότερο χρόνο, περιγράφεται αναλυτικά στο [11] και απαιτεί τον υπολογισμό της ποσότητας:

$$KL(P||Q) \leq \sum_{k=1}^{K} v_k(KL(a_k||\hat{a_k}) + KL(b_k||\hat{b_k}))$$

όπου v^T είναι ένα διάνυσμα στάσιμης κατανομής (stationary distribution) $v^T A = v^T$ και

$$\lim_{n\to\infty} \pi^T A^n = v^T$$

Ο υπολογισμός των στάσιμων αυτών κατανομών, μπορεί να γίνει με τη λύση του αντίστοιχου συστήματος ή τον υπολογισμό του αριστερού ιδιοδιανύσματος του πίνακα που A.

Για τον υπολογισμό του κέντρου κάθε ομάδες (cluster) θα αξιοποιήσουμε την απλούστερη τεχνική της λήψης του μέσου όρου των στοιχείων που εξετάζουμε. Η τιμή αυτή ελαχιστοποιεί την απόσταση Bregman divergence, όπως ειπώθηκε και κατά το θεωρητικό υπόβαθρο και αποδείχτηκε από το Banerjee [3].

Η ποσότητα *log_likelihood* σε αυτήν την περίπτωση υπολογίζεται ως εξής:

$$log(\sum_{k=1}^{K}(centroid\_of\_host\_transition\_matrix[previous\_point][k] * f(x, l_k)))$$

Ένα ενδιαφέρον πρόβλημα που αντιμετωπίσαμε είναι ο υπολογισμός της απόστασης Kullback-Leibner $KL(P||Q)$ στα σημεία που η ποσότητα Q είναι μηδενική, ενώ η ποσότητα P δεν είναι. Ποιοτικά το πρόβλημα αυτό μπορούμε να το αναλογιστούμε ως εξής. Η απόσταση Kullback-Leibler μπορεί να ερμηνευτεί ως το επιπλέον μήκος που χρειάζεται για να σταλθεί ένα μήνυμα από την κατανομή P χρησιμοποιώντας κωδικοποίηση από την κατανομή Q. Αν σε έναν όρο από την κατανομή Q ανατίθεται πολύ μικρή πιθανότητα εμφάνισης, τότε από τον τρόπο κωδικοποίησης θα απαιτείται μεγάλο μήκος χαρακτήρων για την κωδικοποίηση αυτού, εφόσον αναμένεται η εμφάνισή του τόσο σπάνια. Αν αντιθέτως εμφανίζεται συχνά στην κατανομή P, τότε αναγκαστικά το μήνυμα της κατανομής P θα απαιτεί μεγαλύτερο αριθμό από χαρακτήρες για να κωδικοποιηθεί. Μπορούμε επομένως να καταλήξουμε στο συμπέρασμα ότι μηδενική πιθανότητα ενός χαρακτήρα στην κατανομή Q, καθιστά αδύνατη την κωδικοποίηση του μηνύματος.

Τέτοιες περιπτώσεις ωστόσο εμφανίζονται στην περίπτωσή μας, καθώς χρήστες μπορούν να μην εμφανίζουν κίνηση σε κάποιες από τις κατανομές. Υπάρχουν τεχνικές αποφυγής των καταστάσεων αυτών.

- Μπορούμε απλά να αγνοήσουμε τέτοιες περιπτώσεις στο σύνολο της απόστασης δύο πινάκων.

- Μπορούμε να υπολογίσουμε την ελαφρώς διαφορετική απόσταση μεταξύ των κατανομών $P||(P+Q)/2$ και $Q||(P+Q)/2$.

- Μπορούμε να κάνουν μια ομαλοποίηση των κατανομών (smoothing). Πιο συγκεκριμένα αρχικοποιούμε τον πίνακα σε κάποια τιμή, με την αρχικοποίηση να φθίνει αντιστρόφως ανάλογα με την έλευση νέων στοιχείων. Η αρχικοποίηση αυτή μπορεί να είναι τυχαία. Ωστόσο, παρατηρείται η τάση κάθε χρήστης να παραμένει στην ίδια κατάσταση (κατανομή) στην οποία και βρίσκεται με μεγαλύτερη πιθανότητα. Πιο συγκεκριμένα το 70% των παρατηρήσεων εμφανίζονται στην ίδια κατανομή με την προηγούμενη του αντίστοιχου χρήστη. Επομένως μπορούμε να ακολουθήσουμε την τάση αυτή, ορίζοντας ως αρχικό πίνακα τον ακόλουθο:

Καταστάσεις	Μεταβάσεις			
	Καμία (στατικές)	Συγκεντρωτικά	Ομαδοποιημένες	Ανά χρήστη
Συγκεντρωτικά	A παράμετροι: P	-	-	-
Ομαδοποιημένες	B παράμετροι: $K \times P$	D παράμετροι: $K \times P + K \times K$	F παράμετροι: $K \times P + L \times K \times K$	E παράμετροι: $K \times P + N \times K \times K$
Ανά χρήστη	C παράμετροι: $N \times K\_hosts \times P$	Οι καταστάσεις δεν είναι συγκρίσιμες. Είναι αδύνατος ο προσδιορισμός μοντέλου που να καταφέρνει να αποδώσει τα διαφορετικά μοτίβα συμπεριφοράς.		

N: αριθμός χρηστών
M: αριθμός εποχών
P: διασταση δεδομένων
K· αριθμός κατανομών
L: αριθμός ομάδων (clusters)

Πίνακας 4.3. Τα διαφορετικά μοντέλα και ο αριθμός των παραμέτρων που απαιτούν αυτά.

$$transition\_matrix = \begin{bmatrix} \frac{2}{K+1} & \frac{1}{K+1} & \cdots & \frac{1}{K+1} \\ \frac{1}{K+1} & \frac{2}{K+1} & \cdots & \frac{1}{K+1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{K+1} & \frac{1}{K+1} & \cdots & \frac{2}{K+1} \end{bmatrix}$$

Εισάγουμε τεχνητά bias, που ωστόσο εικάζουμε ότι επιφέρει καλύτερα αποτελέσματα από μία τυχαία αρχικοποίηση.

Συγκεντρωτικά οι μέθοδοι και ο αριθμός των παραμέτρων που χρησιμοποιεί η κάθε μία, μπορούν να απεικονιστούν στον Πίνακα 4.3.

ΚΕΦΑΛΑΙΟ 5

Αξιολόγηση αποτελεσμάτων

Οι παραπάνω τεχνικές, όπως αυτές αναλύθηκαν, εφαρμόστηκαν στα δεδομένα του Los Alamos National Laboratory. Τα δεδομένα στο σύνολο αυτό δεν είναι επισημασμένα. Για το σκοπό αυτό θα δημιουργήσουμε ένα τεχνητό σύνολο με σκοπό την ανάδειξη των πλεονεκτημάτων της τελικής μεθόδου που επιλέξαμε σε σχέση με τις υπόλοιπες.

5.1 Τεχνητό σύνολο δεδομένων

Θα δημιουργήσουμε τεχνητά τις κλάσεις από χρήστες-συσκευές που θα παρουσιάζουν παρόμοιες συμπεριφορές. Πιο συγκεκριμένα:

Ορίζουμε ένα σύνολο από κατανομές, από τις οποίες και θεωρούμε ότι προέρχονται οι τιμές-δεδομένα κάθε χρήστη. Στη συνέχεια ορίζουμε τις κλάσεις στις οποίες θα ανήκει κάθε χρήστης. Κάθε κλάση χαρακτηρίζεται από ένα συγκεκριμένο πίνακα μεταβάσεων. Ο πίνακας αυτός παράγεται τυχαία.

Για να ανταποκρίνονται οι συνθήκες των πειραμάτων κατά το δυνατό σε πραγματικές συνθήκες, πραγματοποιούμε τα ακόλουθα:

- Οι κατανομές των διαφορετικών τιμών-δεδομένων επιλέγονται με τέτοιο τρόπο, ώστε να υπάρχει κάποια επικάλυψη μεταξύ αυτών. Για την τιμή x_i, η υπόθεση ότι προήλθε από την κατανομή στην οποία παρουσιάζει τη μεγαλύτερη πυκνότητα μάζας πιθανότητας, θα ισχύει για την πλειονότητα των περιπτώσεων, αλλά όχι για όλες. Αν ορίσουμε ότι k_i είναι η κατανομή από την οποία προήλθε η τιμή x_i και ως $\widehat{k_i}$ την πρόβλεψη από τη διαδικασία αυτή:

$$\widehat{k_i} = \underset{k}{\operatorname{argmax}} f(x_i, l_k)$$

όπου

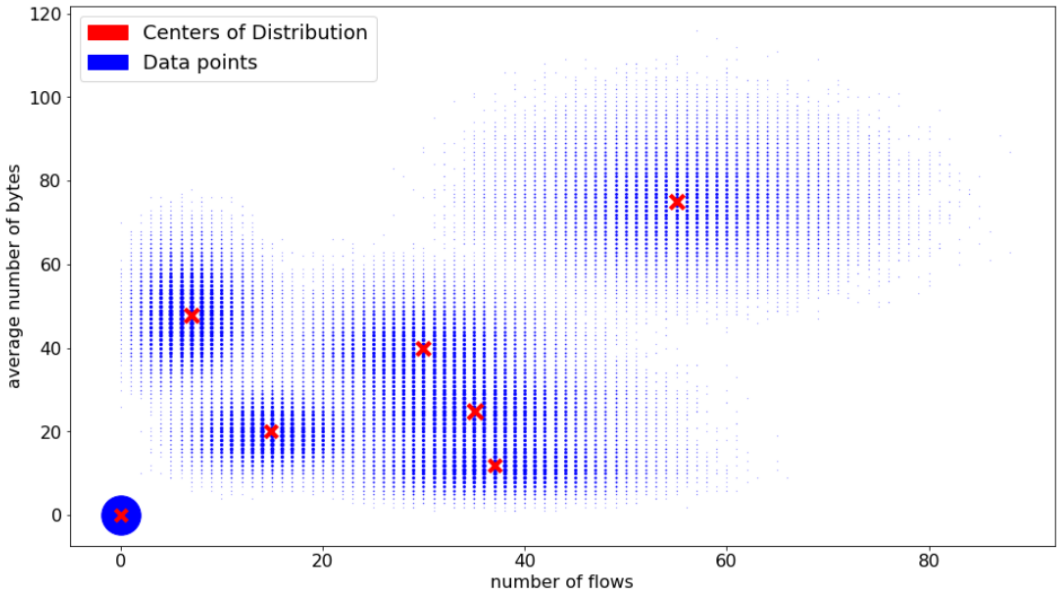
$$f(x_i, l_k) = \prod_{j=1}^{J} \frac{(l_{kj})^{x_{ij}} e^{-l_{kj}}}{x_{ij}!}$$

- Δημιουργούμε κλάσεις από χρήστες, κάθε μία από τις οποίες χαρακτηρίζεται από διαφορετικό αριθμό από χρήστες.

Κάθε χρήστης ξεκινά από μία εκ των καταστάσεων που έχουν οριστεί και μεταβαίνει τυχαία σε κάποια από τις επόμενες σύμφωνα με ένα μηχανισμό ρουλέτας και με βάση την κλάση στην οποία ανήκει. Η επιλογή κάθε τιμής κάθε χρήστη σε μία κατάσταση γίνεται με τυχαίο τρόπο και με βάση το κέντρο της κατανομής.

Δημιουργούμε μια σειρά από δείγματα μέσω της παραπάνω διαδικασίας και εκτελούμε τους αλγορίθμους onlinEM και στη συνέχεια Kullback - Leibler k-mneas.

Στο Σχήμα 5.1, βλέπουμε τις τιμές που προσομοιώσαμε.



Σχήμα 5.1. Τεχνητό σύνολο δεδομένων και τα κέντρα των κατανομών προέλευσης αυτών.

Συγκρίνουμε τις μεθόδους ¨Ομαδοποίηση δεδομένων και μίξη κατανομών' και ¨Ομάδες κρυφών Μαρκοβιανών μοντέλων με ομαδοποίηση δεδομένων¨. Στην πρώτη περίπτωση εξετάζουμε τον αριθμό των τιμών κάθε χρήστη σε κάθε κατανομή. Για το σκοπό αυτό θα χρησιμοποιήσουμε τον αλγόριθμο kMeans της βιβλιοθήκης sklearn της python. Στην δεύτερη περίπτωση, όπως αναλύσαμε στο προηγούμενο κεφάλαιο, ταξινομούμε χρήστες με βάση παρόμοιες μεταβάσεις.

Για την αξιολόγηση των μεθόδων θα χρησιμοποιήσουμε μια σειρά από μετρικές. Ο λόγος για αυτή την επιλογή, είναι ότι δεν υπάρχει μια μοναδική ποσότητα που να

απεικονίζει την ποιότητα των τελικών ομάδων ολοκληρωτικά.

Κλασσική μετρική για τον σκοπό αυτό είναι η normalized mutual info score [27], η οποία ορίζεται ως εξής:

$$NMI(\Omega, C) = \frac{I(\Omega; C)}{\sqrt{H(\Omega) \times H(C)}}$$

όπου Ω είναι οι ομάδες (clusters), C είναι οι αρχικές κλάσεις, $H(.)$ είναι η εντροπία και $I(\omega; C)$ είναι η από κοινού πληροφορία (mutual information) μεταξύ των Ω και C. Θα είναι:

$$\begin{aligned} I(\Omega; C) &= \sum_k \sum_j P(\omega_k \cap c_j) \log \frac{P(\omega_k \cap c_j)}{P(\omega_k) P(c_j)} \\ &= \sum_k \sum_j \frac{|\omega_k \cap c_j|}{N} \log \frac{N|\omega_k \cap c_j|}{|\omega_k||c_j|} \end{aligned}$$

Η ποσότητα αυτή μετρά την ποσότητα της πληροφορίας που κερδίζουμε γνωρίζοντας σε ποια ομάδα ανήκει κάθε τιμή και παίρνει τιμές στο διάστημα [0,1].

Άλλες αποτελούν η rand index [30], η οποία υπολογίζει την ακόλουθη ποσότητα:

$$RI = \frac{True\_Positive + True\_Negative}{True\_Positive + False\_Positive + False\_Negative + True\_Negative}$$

υπολογίζοντας τα ζευγάρια τιμών που κατηγοριοποιήθηκαν σε σωστή ομάδα ή όχι, αναλόγως σε ποια κλάση ανήκουν. Η μετρική homogeneity_score, ελέγχει κατά πόσο κάθε ομάδα περιέχει τιμές της ίδιας κλάσης, ενώ η completeness_score, κατά πόσο όλες οι τιμές της ίδιας ομάδας τοποθετούνται στην ίδια ομάδα.

Στον Πίνακα 5.1 βλέπουμε τα τελικά αποτελέσματα για μια σειρά συγκρίσεων. Τα αποτελέσματα αυτά προέκυψαν παίρνοντας το μέσο όρο 3 επαναλήψεων για κάθε περίπτωση.

Είναι εμφανής η αισθητή διαφορά μεταξύ των δύο μεθόδων. Ειδικότερα, καθώς ο αριθμός των κλάσεων αυξάνεται και αντίστοιχα ο αριθμός χρηστών σε ορισμένες από τις κλάσεις αυτές, μεγαλώνει και η ψαλίδα της διαφοράς μεταξύ των μεθόδων αυτών. Το γεγονός αυτό οφείλεται στη μεγαλύτερη πιθανότητα χρήστες από διαφορετικές κλάσεις να έχουν παρόμοιο αριθμό από τιμές στις κατανομές, ασχέτως από τις μεταβάσεις μέσω των οποίων επιτεύχθηκαν αυτές. Βασιζόμενοι στο γεγονός ότι μεγαλύτερη σημασία στην κατηγοριοποίηση αλλά και στην ανίχνευση ανωμαλιών έχει το στοιχείο αυτό, δηλαδή οι μεταβάσεις μεταξύ των κατανομών και όχι ο

Αριθμός χρηστών ανά κλάση	Αλγόριθμος	NMI	RI	HG	CMP
[30, 30, 30, 30, 30]	B	1.0	1.0	1.0	1.0
	F	0.984	0.983	0.984	0.984
[15, 20, 25, 30, 35, 40, 45, 50]	B	0.990	0.984	0.985	0.995
	F	0.805	0.698	0.808	0.803
[10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 80, 150]	B	0.983	0.951	0.989	0.977
	F	0.849	0.707	0.878	0.822
[8, 12, 15, 18, 20, 25, 33, 36, 40, 50, 55, 60, 65, 150, 250, 450]	B	0.963	0.898	0.984	0.944
	F	0.749	0.479	0.831	0.676

NMI: normalized mutual information
RI: rand index
HG: homogeneity score
CMP: completeness score
Αλγόριθμος B: mixture of Poisson
Αλγόριθμος F: mixture of Poisson and clustering of transition matrices

Πίνακας 5.1. Ποιοτική σύγκριση των ομάδων από τις δύο μεθόδους.

αριθμός στοιχείων ανά κατανομή, καταλήγουμε στο συμπέρασμα ότι η μέθοδός μας είναι καταλληλότερη για το σκοπό αυτό.

5.2 Πραγματικά δεδομένα

Στη συνέχεια εργαζόμαστε με τα δεδομένα του Los Alamos National Laboratory. Στο προηγούμενο παράδειγμα είχαμε την πολυτέλεια να προσδιορίσουμε ορισμένες παραμέτρους εύκολα, καθώς γνωρίζαμε τις κατανομές από τις οποίες προέκυψαν τα δεδομένα που δημιουργήσαμε. Στην περίπτωση αυτή δεν έχουμε αυτή την πολυτέλεια, γεγονός που καθιστά ιδιαίτερα σημαντική την επιλογή των σωστών παραμέτρων αυτών.

Πιο σημαντική παράμετρος είναι ο αριθμός των κατανομών που θα χρησιμοποιηθούν για την αρχική απεικόνιση των δεδομένων. Προφανώς η επιλογή μεγαλύτερου αριθμού κατανομών, θα οδηγήσει σε πιο ακριβή αναπαράσταση όλων των σημείων. Ωστόσο, η αύξηση των κατανομών οδηγεί και σε αύξηση του αριθμού παραμέτρων προς υπολογισμό και ως συνέπεια του υπολογιστικού κόστους. Θα πρέπει επομένως να υπάρχει ένας συμβιβασμός στον αριθμό των παραμέτρων και στο κόστος αυτό. Λύση σε αυτό το πρόβλημα είναι η προσθήκη στο όρο της πιθανοφάνειας (likelihood) και ενός όρου που σχετίζεται με τον αριθμό των παραμέτρων αυτών. Ένας τέτοιος δείκτης είναι ο Bayessian Information Criterion [34]. Ορίζεται δε ως εξής:

$$BIC = -2\log(L_n(\theta^*)) + \log(n)N_\theta$$

όπου $L_n(\theta^*)$ είναι η καλύτερη δυνατή πιθανοφάνεια και N_θ είναι ο αριθμός των ανεξαρτήτων παραμέτρων. Στην περίπτωση του μοντέλου μας ελεύθεροι παράμετροι είναι τα κέντρα των κατανομών $lambdas$ και τα βάρη αυτών (a priori πιθανότητες) $gammas$, με την ιδιαιτερότητα ότι ισχύει η σχέση

$$\sum_{i=1}^{K} gamma_i = 1$$

Επομένως τελικά ο αριθμός των ελεύθερων παραμέτρων είναι:

$$(J+1) * K - 1$$

όπου J είναι ο αριθμός των χαρακτηριστικών που εξετάζουμε και K ο καθαυτός αριθμός των κατανομών. Επίσης στην περίπτωση μας που εξετάζουμε online αλγόριθμο, η πιθανοφάνεια αντικαθίσταται από την ποσότητα sample log likelihood [26]. Τελικά:

$$BIC^* = -\frac{2}{M}\sum_{m=1}^{M} \log(L_N(X_m, \theta^{(m)}) + \log(N)N_\theta$$

όπου M είναι ο αριθμός των εποχών που εξετάζουμε και X_m είναι τα δεδομένα που σχετίζονται με μία εποχή. Εξετάζουμε τα αποτελέσματα για μια σειρά αριθμού κατανομών. Τα αποτελέσματα φαίνονται στον πίνακα 5.2. Η καλύτερη παράμετρος είναι για $K = 7$, με την οποία και θα συνεχίσουμε στη συνέχεια.

Αριθμός κατανομών	4	5	6	7	8	9	10
AVG log-likelihood	3947.3	3874.3	3836.5	-3756.4	-3752.6	-3748.1	-3741.8
BIC	7979.6	7856.7	7804.4	7667.2	7682.9	7697.1	7707.7

BIC: Bayesian Information Criterion

Πίνακας 5.2. Επιλογή αριθμού κατανομών.

Έχοντας ορίσει τον αριθμό των κατανομών που αντιπροσωπεύουν τα αρχικά δεδομένα, θα πρέπει στη συνέχεια στα προχωρήσουμε στον ορισμό του κατάλληλου αριθμού ομάδων που θα δημιουργηθούν από την τελική ομαδοποίηση των πινάκων μεταβάσεων, που θα αντιπροσωπεύουν και τα τελικά δυνατά προφίλ κάθε χρήστη.

Μία από τις πιο απλές αλλά ταυτόχρονα και ευρέως χρησιμοποιούμενες μετρικές για το σκοπό αυτό, είναι η διασπορά της κάθε ομάδας (cluster). Για το cluster C η ποσότητα αυτή θα είναι:

$$I_C = \sum_{i \in C} distance(tm_i, tm_C)$$

όπου tm_i είναι ο πίνακας μετάβασης του χρήστη i που ανήκει στο cluster C με κέντρο tm_C. Με τον παραπάνω τρόπο υπολογίζουμε την inertia για το δεδομένο clustering ως εξής:

$$I = \sum_{C=1}^{K} I_C$$

Επίσης θα χρησιμοποιήσουμε τη μετρική silhouette coefficient, που δίνει τη δυνατότητα υπολογισμού της συνοχής μεταξύ των clusters. Συγκρίνει τις αποστάσεις του κάθε σημείου σε σχέση με τα άλλα σημεία του cluster στο οποίο ανήκει, σε σχέση με τις αποστάσεις του από σημεία άλλων clusters [31]. Πιο συγκεκριμένα:

$$Silhouette = \frac{1}{N} \sum_{i=1}^{N} s(i)$$

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

όπου $a(i)$ η μέση απόσταση της τιμής i από τις υπόλοιπες τιμές στο ίδιο cluster, ενώ $b(i)$ η μικρότερη μέση απόσταση από τιμές που ανήκουν σε διαφορετικά clusters. Αν ορίσουμε ως:

$$d(i, C) = average\, dissimilarity\, of\, i\, to\, all\, objects\, of\, C$$

και A το cluster στο οποίο ανήκει το στοιχείο i τότε:

$$b(i) = \underset{C \neq A}{\text{minimum}}\, d(i, C)$$

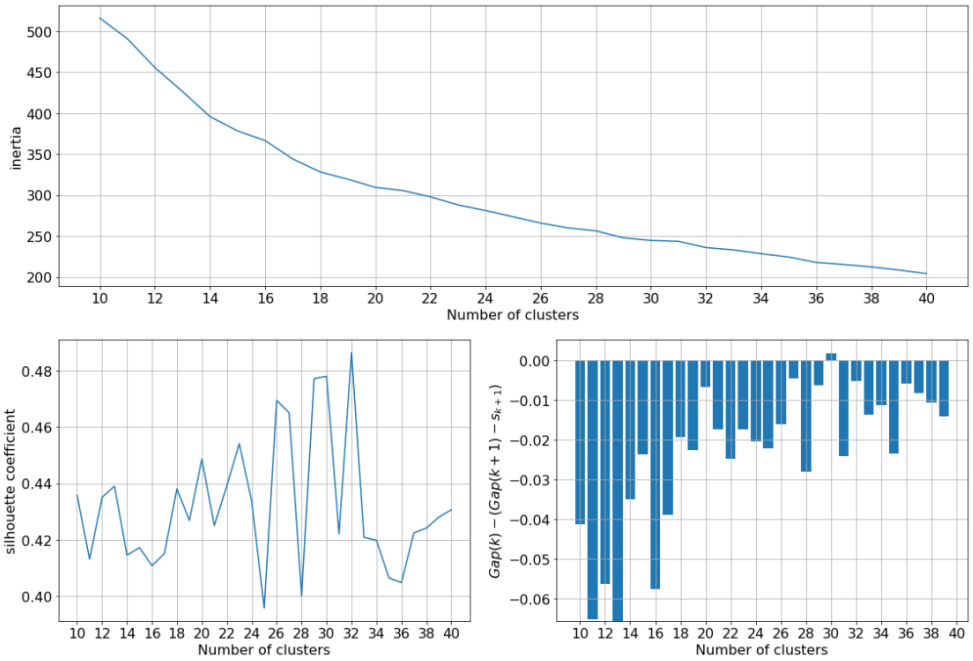
Τέλος η gap statistic [38], αναλώνεται στον ακόλουθο υπολογισμό:

$$Gap_n(k) = E_n^*\{\log W_k^*\} - \log W_k$$

Συγκρίνεται το αποτέλεσμα της διαδικασίας του clustering στα τελικά δεδομένα, σε σχέση με την ίδια διαδικασία σε δεδομένα που δεν παρουσιάζουν προφανείς ομάδες. Για τον υπολογισμό της $E_n^*\{\log W_k^*\}$, παράγουμε B τυχαία δείγματα μέσω δειγματοληψίας Monte Carlo. Τελικό επιλέγεται ο ιδανικός αριθμός από clusters k, ως το μικρότερο k, για το οποίο ισχύει ότι $Gap(k) \geq Gap(k+1) - s_{k+1}$, όπου $s_k = \sqrt{1 + \frac{1}{B}} sd(k)$ και $sd(k)$ η τυπική απόκλιση των $\log W_k^*$.

Δοκιμάζουμε για ένα σύνολο διαφορετικού αριθμού clusters όπως φαίνεται στο Σχήμα 5.2. Για κάθε αριθμό clusters τρέχουμε 10 φορές τον παραπάνω αλγόριθμο, κρατώντας την περίπτωση που οδηγεί στο χαμηλότερο inertia score. Όπως και

στον online Expectation Maximization αλγόριθμο, η αρχικοποίηση των κέντρων γίνεται μέσω παραλλαγής του αλγορίθμου k-means ++, ώστε να γίνεται χρήση της κατάλληλης απόστασης, για τη σύγκριση δύο πινάκων μεταβάσεων. Ύστερα από πειραματικές δοκιμές, η μέθοδος οδηγεί συστηματικά σε καλύτερα τοπικά ελάχιστα. Χρησιμοποιούμε συνδυασμό των μετρικών για την επιλογή της κατάλληλης τιμής. Στην περίπτωση της inertia μπορούμε να χρησιμοποιήσουμε το εμπειρικό κριτήριο του elbow. Επιλέγουμε τη χαμηλότερη τιμή με τα καλύτερα αποτελέσματα για $k = 30$.

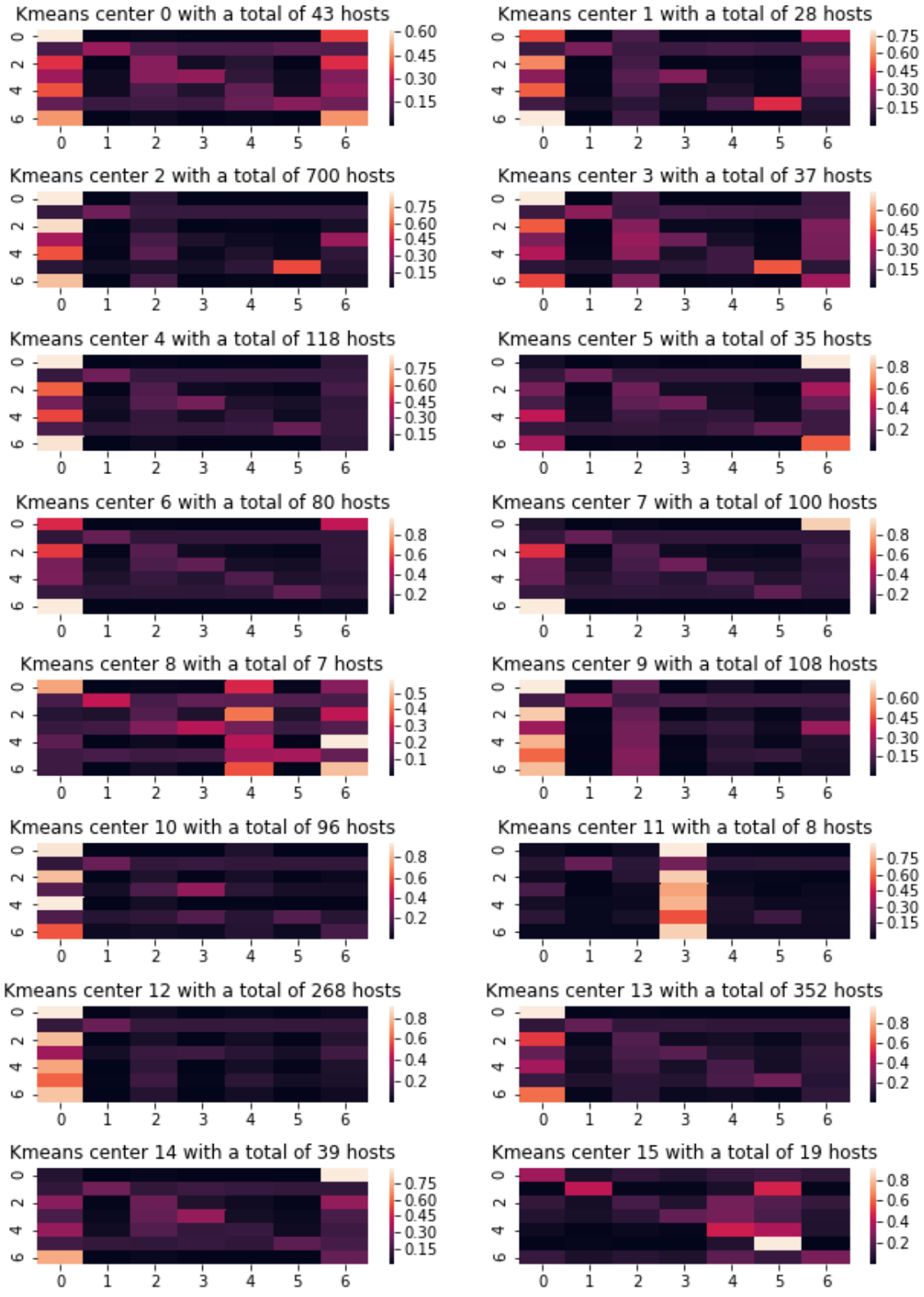


Σχήμα 5.2. Επιλογή αριθμού προφίλ χρηστών.

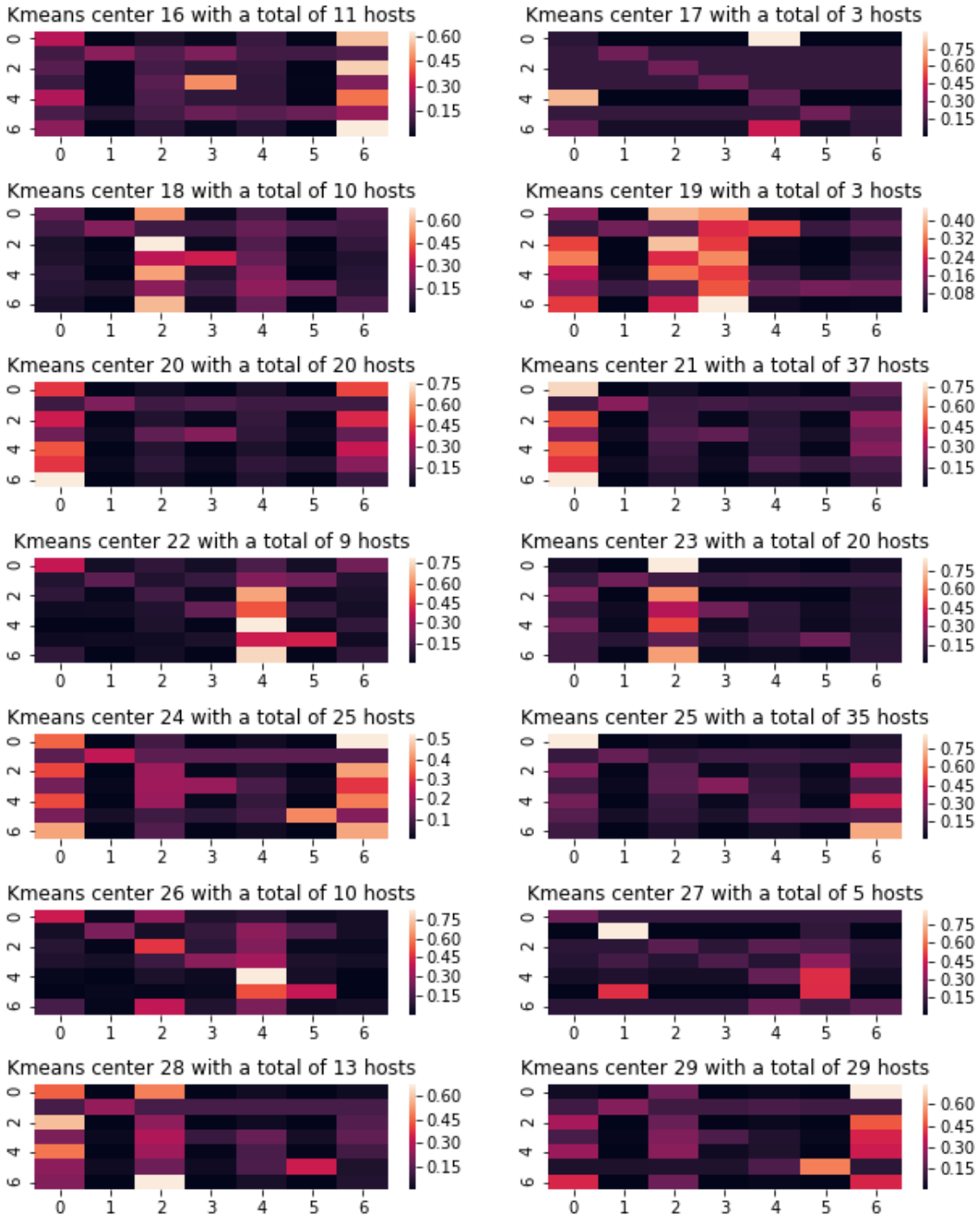
Από τα παραπάνω καταλήγουμε τελικά σε έναν αριθμό προφίλ για τους χρήστες που εξετάζουμε. Τα προφίλ αυτά χαρακτηρίζουν τις δυνατές μεταβάσεις που μπορεί να πραγματοποιήσει ο κάθε χρήστης. Τα προφίλ αυτά φαίνονται στα Σχήματα 5.3 και 5.4.

Η βελτίωση που επιτυγχάνεται από την παραπάνω διαδικασία μπορεί να αποτυπωθεί και στην ποσότητα *log_likelihood*. Πιο συγκεκριμένα στον Πίνακα 5.3. Ο τρόπος υπολογισμού αυτής παρουσιάστηκε στο προηγούμενο κεφάλαιο.

Τέλος μπορούμε να εφαρμόσουμε ένα συνδυασμό του πίνακα μεταβάσεων του κάθε χρήστη με αυτό του cluster στο οποίο αυτός ανήκει. Θα έχουμε:



Σχήμα 5.3. Κέντρα 0-15 KL-kmeans αλγορίθμου.



Σχήμα 5.4. Κέντρα 16-29 KL-kmeans αλγορίθμου.

$$centroid\_of\_host\_transition\_matrix[previous\_point] \times \lambda +$$
$$host\_specific\_transition\_matrix[previous\_state] \times (1 - \lambda)$$

Όπως και στις περισσότερες περιπτώσεις στη στατιστική η τεχνική αυτή βελτιώνει περαιτέρω τα αποτελέσματα. Η περίπτωση αυτή αναφέρεται ως Mixture of Poissons - average of cluster's and host's transition matrix στον Πίνακα 5.3.

Αλγόριθμος	log-likelihood
Mixture of Poissons	-1.6940
Mixture of Poissons - global transition matrix	-1.5784
Mixture of Poissons - clusters of transition matrices	-1.5183
Mixture of Poissons - average of cluster's and host's transition matrix	-1.4910

Πίνακας 5.3. Αποτελέσματα negative log-likelihood.

Αξίζει επίσης να σημειώσουμε για άλλη μία φορά τη σημασία των μεταβάσεων μεταξύ των διαφορετικών καταστάσεων. Για να αναδείξουμε το γεγονός αυτό, προβαίνουμε στην εξής διαδικασία. Δημιουργούμε χρήστες, οι οποίοι παρουσιάζουν την ίδια ακριβώς συνολικά κίνηση σε σχέση με τους υπάρχοντες, αλλά με διαφορετική χρονική σειρά. Με τον τρόπο αυτό επιδιώκουμε να υπολογίσουμε σε ποιο βαθμό επηρεάζονται τα αποτελέσματα της δημιουργίας των προφίλ. Στον Πίνακα 5.4 βλέπουμε τα αποτελέσματα αυτής. Όπως παρατηρούμε, η δημιουργία των τυχαίων μεταβάσεων αυτών προκαλεί μια πολύ μεγάλη άνοδο των αποστάσεων των νέων πινάκων μετάβασης σε σχέση με κέντρα του αλγορίθμου kl-k-means στα οποία άνηκε ο προηγούμενος χρήστης. Η ίδια άυξηση αποτυπώνεται λαμβάνοντας υπόψη και διαφορεικά κέντρα, τα οποία πιθανώς βρίσκονται πιο κοντά στον νέο χρήστη. Σε αυτήν την περίπτωση η επανάληψη της διαδικασίας του clustering, δεν επιφέρει ουσιαστικά αποτελέσματα. Αυτό διότι στη μέση περίπτωση, από τη τυχαία επιλογή των σημείων, ο πίνακας μεταβάσεων θα έχει τις ίδιες τιμές για κάθε κατάσταση, αυτήν του ποσοστού συμμετοχής σε κάθε κατανομή κάθε χρήστη. Αν μετρήσουμε τα σημεία σε κάθε κατάσταση ενός χρήστη i, αυτά θα είναι ένας πίνακας μεγέθους K: $Points\_per\_cluster_i = [p_1 p_2 \dots p_K]$. Τότε ο πίνακας μετάβασης του νέου χρήστη θα συγκλίνει στην τιμή: $Transition\_matrix_i[j] = \frac{1}{sum(Points\_per\_cluster_i)}[p_1 p_2 \dots p_K]$ για κάθε j. Επομένως η διαδικασία εκφυλίζεται σε απλό clustering ανάλογα με τα σημεία ανά κατανομή.

Πέρα από τον αλγόριθμο k-means, μπορούμε επίσης να αξιοποιήσουμε τεχνικές

Ποσοστό αύξησης απόστασης σε σχέση με το κέντρο στο οποίο άνηκε ο χρήστης	149.80%
Ποσοστό αύξησης απόστασης σε σχέση με το νέο κοντινότερο κέντρο	40.59%

Πίνακας 5.4. Δημιουργούμε χρήστες με την ίδια συνολική κίνηση σε σχέση με τους υπάρχοντες, αλλά σε τυχαίες μεταξύ τους χρονικές στιγμές. Παρατηρούμε σημαντικά ποσοστά αύξησης των αποστάσεων από τα κοντινότερα κέντρα του kl-k-means.

που βασίζονται σε σχετικές πυκνότητες δεδομένων, με βάση την ίδια μετρική απόστασης που ήδη χρησιμοποιήσαμε. Μια τέτοια ποσότητα είναι ο δείκτης Local Outlier Factor [6]. Ο αλγόριθμος αυτός παρουσιάζει μεγάλη δημοτικότητα στην περιοχή της ανίχνευσης ανωμαλιών, ενώ θα μελετηθεί σύντομα στην επόμενη υποενότητα.

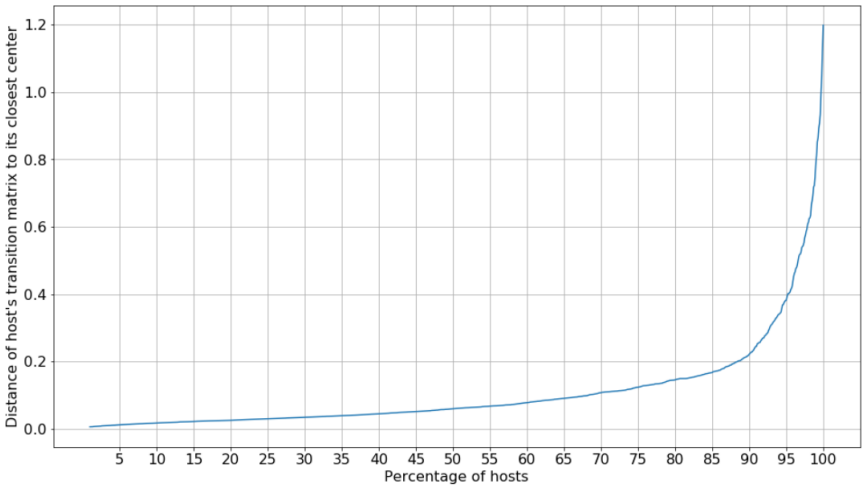
5.3 Ανίχνευση ανωμαλιών

Εφαρμόζοντας τις παραπάνω αλγοριθμικές διαδικασίες, τελικά καταλήγουμε και σε ένα σύνολο αποδεκτών συμπεριφορών. Αντιστοίχως, προκύπτει ένας αριθμός χρηστών, που απέχει αισθητά από τα κέντρα των clusters στο οποίο και αυτοί έχουν ανατεθεί. Με αυτόν τον τρόπο, μπορούμε να προσδιορίσουμε χρήστες με αποκλίνουσες συμπεριφορές, των οποίων η κίνηση αξίζει περαιτέρω διερεύνηση.

5.3.1 Χρήστες με ιδιάζουσα συμπεριφορά

Από τη δημιουργία των παραπάνω προφίλ μπορούμε νε εξάγουμε μιας μορφής score, σχετικά με την απόσταση κάθε χρήστη από το κοντινότερο σε αυτόν προφίλ. Αναπόφευκτα αυτόματα, μπορούμε να αναγνωρίσουμε κάποιους χρήστες των οποίων η συμπεριφορά τείνει να διαφοροποιηθεί σε σχέση με αυτήν των υπολοίπων. Όπως μπορούμε να παρατηρήσουμε και από το Σχήμα 5.5, το μεγαλύτερο ποσοστό των χρηστών βρίσκεται σε κοντινή απόσταση από το κέντρο-προφίλ που τους χαρακτηρίζει, ενώ ένα μικρό ποσοστό αυτών σε μεγαλύτερη απόσταση. Τους χρήστες αυτούς θα εξετάσουμε σύντομα στη συνέχεια. Προτού συνεχίσουμε, αξίζει να σημειώσουμε ότι αλλάζοντας τον αριθμό των προφίλ, η πλειονότητα των χρηστών με ιδιάζουσα συμπεριφορά παραμένει η ίδια. Δηλαδή ως επί των πλείστων, οι ίδιοι χρήστες εμφανίζονται να παρουσιάζουν τις πιο ιδιάζουσες συμπεριφορές. Αυξάνοντας περαιτέρω αυτά σε μεγαλύτερο αριθμό, αναμένουμε πιθανώς διαφορετικά

αποτελέσματα.



Σχήμα 5.5. Άνω κατώφλι της απόστασης κάθε χρήστη από το κοντινότερο σε αυτόν κέντρο, σε συνάρτηση με το ποσοστό των χρηστών.

Μπορούμε να κατηγοριοποιήσουμε τους χρήστες αυτούς με την ιδιαίτερη συμπεριφορά σε δύο μεγάλες κατηγορίες, ανάλογα με τους λόγος που οδηγούν στην ιδιαίτερη αυτή συμπεριφορά.

- Στην πρώτη κατηγορία μπορούμε να κατατάξουμε χρήστες των οποίων οι διαδικτυακές κινήσεις μπορούν να θεωρηθούν φυσιολογικές, αλλά οι μεταβάσεις μεταξύ των καταστάσεων αυτών όχι.

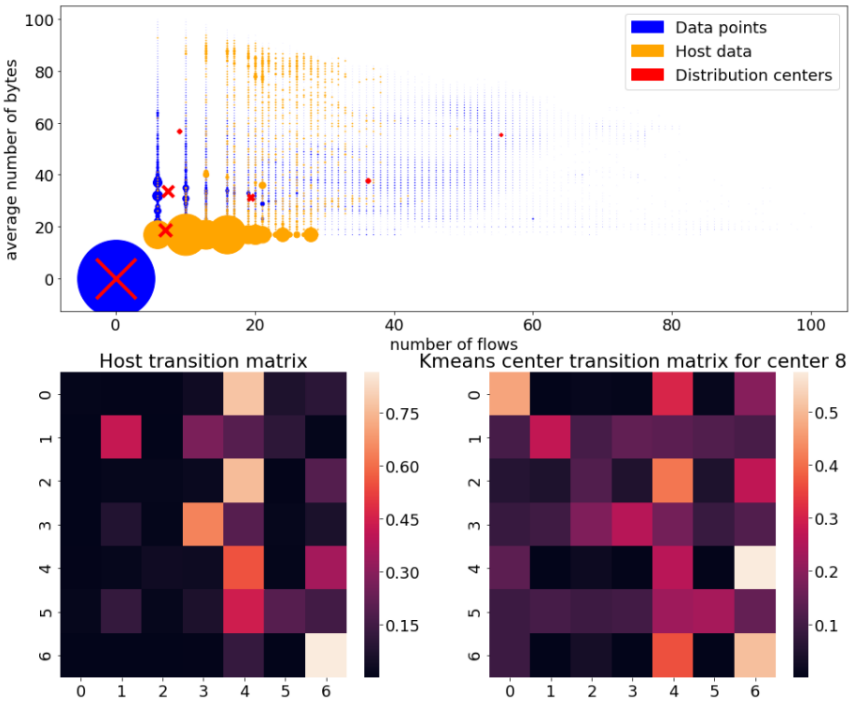
- Υπάρχουν βεβαίως και χρήστες των οποίων η κίνηση είναι δύσκολο να αντιπροσωπευτεί ικανοποιητικά από τον αριθμό των κατανομών που έχουμε επιλέξει. Ορισμένες στιγμές κίνησης, μικρές στο πλήθος, εντοπίζονται σε σημεία μακριά από τις κατανομές, όπως αυτές έχουν προσαρμοστεί.

Στην περίπτωσή μας έχουμε ασχοληθεί με την αναγνώριση ιδιαιτεροτήτων μεταξύ της γενικότερης κίνησης των χρηστών. Η αναγνώριση έκτοπων τιμών για μια δεδομένη χρονική στιγμή, αποτελεί σημαντικό πρόβλημα, αλλά ένα που δε θα μας απασχολήσει σε αυτή την εργασία.

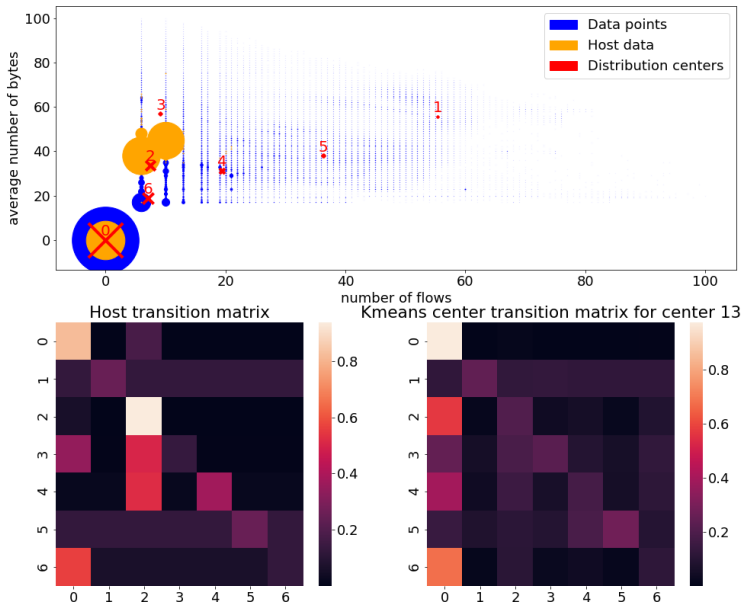
Ακολουθούν ορισμένα παραδείγματα χρηστών με ιδιάζουσα συμπεριφορά στα Σχήματα 5.6, 5.7 και 5.8.

Αξίζει να κάνουμε ορισμένες παρατηρήσεις για τις κινήσεις αυτών.

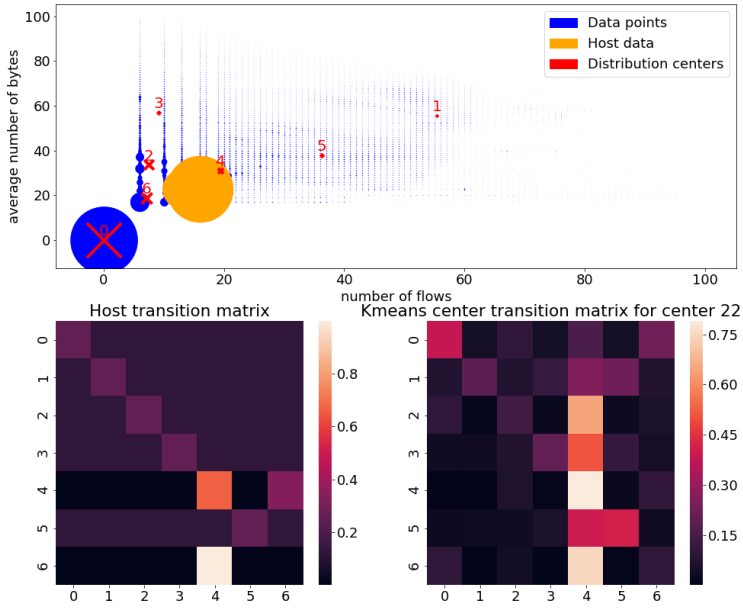
- Όπως προαναφέρομαι οι περισσότεροι χρήστες χαρακτηρίζονται από μηδενική κίνηση την πλειονότητα των χρονικών διαστημάτων που εξετάζουμε. Κάτι



Σχήμα 5.6. Κίνηση χρήστη C2519.



Σχήμα 5.7. Κίνηση χρήστη C202.



Σχήμα 5.8. Κίνηση χρήστη C1340.

τέτοιο δε συμβαίνει αναγκαστικά για τους χρήστες που εξετάζουμε στο σημείο αυτό.

- Οι χρήστες γενικότερα, αλλά και αυτοί που μας απασχολούν στο σημείο αυτό, κατά κανόνα δεν εμφανίζουν κίνηση σε όλες τις καταστάσεις που έχουμε ορίσει ή αν αυτό συμβαίνει δεν συμβαίνει με την ίδια συχνότητα. Το γεγονός αυτό είναι αναμενόμενο και θεμιτό. Παράλληλα ωστόσο αυτές οι καταστάσεις στις οποίες ο χρήστης εμφανίζει κίνηση με μικρή συχνότητα μπορεί να συμβάλλουν σημαντικά στον υπολογισμό της απόστασης. Ωστόσο προσδοκούμε ότι καθώς στην ίδια ομάδα - προφίλ θα ανήκουν χρήστες με παρόμοια χαρακτηριστικά, οι διαφορές αυτές δε θα είναι σημαντικές. Κάτι τέτοιο σαφώς δε συμβαίνει αναγκαστικά στους χρήστες με ιδιάζουσα συμπεριφορά που παρουσιάζουμε.

Μπορούμε επίσης να μελετήσουμε μια διαφορετική μέθοδο, αυτή του Local Outlier Factor [6]. Αρνητικό της μεθόδου αυτής είναι ότι απαιτεί τον καθορισμό του αριθμού των γειτόνων με βάση των οποίων υπολογίζουμε αν ένας χρήστης έχει ιδιάζουσα συμπεριφορά, αριθμός ο οποίος παραμένει σταθερός για όλους τους χρήστες. Στην περίπτωσή μας και με βάση τον αριθμό των χρηστών σε κάθε kl-k-means cluster, μπορούμε να υποθέσουμε ότι ο αριθμός 15 ενδείκνυται. Τα αποτελέσματα από αυ-

τή τη μέθοδο δεν είναι τόσο ξεκάθαρα συγκρίνοντάς τα με τα προηγούμενα. Το φαινόμενο αυτό ενδεχομένως οφείλεται στη μεγάλη διάσταση εισόδου. Οι μέθοδοι κοντινότερου γείτονα είναι γνωστό ότι με μη κατάλληλα διαμορφωμένα δεδομένα υποφέρουν από περιπτώσεις υψηλών false positives, ενώ απαιτούν και υψηλό υπολογιστικό κόστος.

5.3.2 Τεχνητοί χρήστες με απροσδόκητη συμπεριφορά

Εξετάζουμε την ακρίβεια του συστήματος αυτού, με την εισαγωγή νέων χρηστών, οι οποίοι εμφανίζουν τυχαία συμπεριφορά, η οποία και επιθυμούμε να αναγνωριστεί. Ο τρόπος δημιουργίας της τυχαίας αυτής συμπεριφοράς ποικίλει. Στη συνέχεια θα εξετάσουμε κάποιες περιπτώσεις. Καθώς οι περιπτώσεις αυτές εμπεριέχουν σε μεγάλο βαθμό τυχαιότητα, οι διαδικασίες θα επαναληφθούν πολλαπλές φορές για την εξαγωγή αξιόπιστων συμπερασμάτων. Οι τελικές μετρήσεις φαίνονται στον Πίνακα 5.5.

- Δημιουργία τυχαίου πίνακα μεταβάσεων. Στην περίπτωση αυτή υποθέτουμε τυχαίες μεταβάσεις μεταξύ των καταστάσεων που έχουμε εμείς ορίσει. Δεν ασχολούμαστε με τις αρχικές μετρήσεις του χρήστη όσον αφορά τις ροές που καταγράφηκαν από σύστημα συλλογής αυτών, αλλά μονάχα με τις μεταβάσεις μεταξύ των ήδη προσδιορισμένων κατανομών Poisson. Σε αυτήν την περίπτωση δεν διαθέτουμε τον αριθμό των σημείων ανά κατάσταση και επομένως η αρχικοποίηση του πίνακα μεταβάσεων, όπως την ορίσαμε παραπάνω, σε συνδυασμό με την εκθετική μείωση αυτής δεν είναι δυνατή. Υποθέτουμε ότι για αρκετά μεγάλο αριθμό δεδομένων δεν επιφέρει σημαντικές αλλαγές.

- Δημιουργία δεδομένων μέσω τυχαίου συνδυασμού χαρακτηριστικών (features), από το σύνολο τιμών. Στην περίπτωση αυτή, εξετάζουμε τα χαρακτηριστικά (number of flows, mean number of bytes) ξεχωριστά. Επιλέγουμε για κάθε ένα ξεχωριστά κάποια τιμή από αυτές που έχουν εμφανιστεί ξανά στο παρελθόν και τις συνδυάζουμε για τη δημιουργία μιας καινούριας μέτρησης ενός χρήστη για μία εποχή.

- Τέλος για να αξιοποιήσουμε τη σχέση μεταξύ των χαρακτηριστικών, δημιουργούμε χρήστες των οποίων η κίνηση κάθε χρονική στιγμή επιλέγεται από το σύνολο των διαφορετικών μετρήσεων που έχουν ήδη εμφανιστεί στο παρελθόν. Με τη μέθοδο αυτή επιδιώκουμε να εξαλείψουμε πιθανώς μη ρεαλιστικές μεταβάσεις κίνησης.

Τρόπος παραγωγής δεδομένων	Μικρότερη απόσταση από κέντρο kl-k-menas	Ποσοστό ανωμαλίας (%)
Τυχαίος πίνακας μεταβάσεων	1.6671	100.0
Τυχαίος συνδυασμός τιμών που έχουν εμφανιστεί	1.0070	99.69
Τυχαίος συνδυασμός κινήσεων παρελθόντος	1.0944	99.73
Τυχαίος συνδυασμός κινήσεων παρελθόντος διατηρώντας τις ίδιες πιθανότητες εμφάνισης	0.1923	87.61

Ποσοστό ανωμαλίας: Το ποσοστό των χρηστών, των οποίων η απόσταση του πίνακα μετάβασης από το κοντινότερό τους κέντρο είναι μικρότερη

Πίνακας 5.5. Αποτελέσματα ανίχνευσης ανωμαλιών.

ΚΕΦΑΛΑΙΟ 6

Επίλογος

6.1 Σύνοψη και συμπεράσματα

Στην παρούσα διπλωματική εργασία παρουσιάστηκε ένας μηχανισμός κατηγοριοποίησης δεδομένων ανά χρήστη-συσκευή και ομαδοποίησης αυτών με βάση ένα Μαρκοβιανό μοντέλο.

Πιο συγκεκριμένα, αναπαριστούμε τις τιμές των χρηστών-συσκευών με τη χρήση μιας μίξης κατανομών (Poisson στην περίπτωσή μας, αλλά η ίδια διαδικασία μπορεί να εφαρμοστεί για διαφορετικό είδος κατανομών, όπως για παράδειγμα Gaussian). Τονίζουμε ότι για να ανταποκριθούμε με τη μεγάλη ποσότητα των δεδομένων και να παραγάγουμε ένα σύστημα ανταποκρίσιμο, η παραπάνω διαδικασία πραγματοποιείται σε πραγματικό χρόνο (online), με τη χρήση ενός online Expectation-Maximization αλγορίθμου [26, 25].

Στη συνέχεια, μπορούμε από τον αρχικό χώρο εισόδου, να περάσουμε σε έναν πολύ χαμηλότερου, ίσο με τον αριθμό των κατανομών που έχουμε χρησιμοποιήσει για την αναπαράσταση αυτών. Η μετάβαση αυτή γίνεται με απόλυτο τρόπο (hard), βασιζόμενοι στην ισχυρή σύνδεση μεταξύ των αλγορίθμων k-means και Expectation Maximization.

Τέλος από τις μεταβάσεις των χρηστών-συσκευών μεταξύ των διαφορετικών καταστάσεων αυτών μπορούμε να δημιουργήσουμε προφίλ διαφορετικής συμπεριφοράς. Μπορούμε στη συνέχεια να κατηγοριοποιήσουμε τους χρήστες και να καταλήξουμε σε χρήσιμες πληροφορίες σχετικά με την κίνησή τους. Η παραπάνω διαδικασία οδηγεί επίσης στην αναγνώριση χρηστών που παρουσιάζουν αποκλίνουσα συμπεριφορά, δηλαδή πρότυπα κίνησης που δε συμπίπτουν με κάποιο από τα αναγνωρισμένα.

Βασικός παράγοντας που οδήγησε στην εκπόνηση της δουλειάς αυτής, αποτελεί ο ολοένα αυξανόμενο αριθμός διαθέσιμων δεδομένων πραγματικού χρόνου και η

δυσκολία που υπάρχει στην αποθήκευση και επεξεργασία της πληροφορίας αυτής. Οι τεχνικές που παρουσιάζουμε αποσκοπούν στη μετατροπή της πληροφορίας και τη σημαντική μείωση στον όγκο της πληροφορίας που απαιτείται να αποθηκευτεί ανά χρήστη-συσκευή. Παράλληλα η τελική μορφή διάθεσης αυτής, παρέχει στο διαχειριστή οποιουδήποτε συστήματος άμεσα χρήσιμες πληροφορίες, οδηγώντας τον προς τη σωστή κατεύθυνση για περαιτέρω διερεύνηση.

6.2 Βελτιώσεις και μελλοντικές επεκτάσεις

Η δουλειά που παρουσιάσαμε αποτελεί τη βάση ενός ενδεχομένως πιο εξελιγμένου συστήματος. Υπάρχουν ορισμένα σημεία που ήταν πέρα από τα πλαίσια της διπλωματικής αυτής και στα οποία δε θέσαμε ιδιαίτερο βάρος.

Όπως προαναφέρθηκε, μεγάλο πλεονέκτημα του αλγορίθμου είναι το γεγονός ότι είναι πραγματικού χρόνου, δηλαδή ανανεώνεται με βάση νέα δεδομένα, καθώς αυτά καταφθάνουν. Δεν ασχοληθήκαμε ωστόσο με δυναμικά φαινόμενα από την άποψη περιπτώσεων που παρουσιάζεται γενικότερη αλλαγή των κατανομών. Το γεγονός αυτό μπορεί να συμβεί σταδιακά με το χρόνο ή σε περιπτώσεις flush-crowd events, όταν παρουσιάζεται απότομη αύξηση της κίνησης. Για παράδειγμα, η οργάνωση μιας εκδήλωσης σε ένα χώρο, μπορεί να οδηγήσει σε απρόσμενη αύξηση της κίνησης στο αντίστοιχο wifi. Ομοίως, μια είδηση μπορεί να οδηγήσει σε αύξηση των αναζητήσεων για τα πρόσωπα που συνδέονται με το συγκεκριμένο συμβάν. Οι τεχνικές αντιμετώπισης του προβλήματος αυτού αναλώνονται κυρίως σε διαδικασίες exponential forgetting ή την επεξεργασία δεδομένων σε χρονικά παράθυρα και τη λήψη τελικών αποτελεσμάτων με βάση έναν αριθμό των τελευταίων παραθύρων αυτών [40, 37]. Οι τεχνικές αυτές θα πρέπει να εφαρμοστούν τόσο στο βήμα της αναβάθμισης των παραμέτρων των κατανομών στον online expectation maximization αλγόριθμο, όσο και στην ανανέωση των πινάκων μεταβάσεων και των αντίστοιχων κέντρων μέσω του αλγορίθμου kl-kmeans.

Σημαντική πρόκληση αποτελεί επίσης η ανίχνευση των ανωμαλιών. Για το σκοπό αυτό θα πρέπει να λαμβάνονται υπόψη τόσο οι σχετικές αποστάσεις των τιμών της κίνησης κάθε νέου χρήστη σε σχέση με τις κατανομές, όσο και η απόσταση του πίνακα μεταβάσεων του χρήστη από το κοντινότερο κέντρο σε αυτόν. Σημαντικό ζήτημα είναι επίσης ο τρόπος με τον οποίο θα γίνεται ταξινόμηση για έναν καινούριο χρήστη. Απαιτείται το πέρας ενός χρονικού διαστήματος για τη συγκέντρωση επαρκών αποτελεσμάτων. Το διάστημα αυτό μπορεί να ποικίλει ανάλογα με το βαθμό στον οποίο μεταβάλλεται η συμπεριφορά του χρήστη. Το φαινόμενο αυτό έχει

αναλυθεί σε μεγαλύτερο βαθμό για τη σωστή κατηγοριοποίηση εφαρμογών σε data centers [9].

Τέλος αναφερθήκαμε κυρίως στο αλγοριθμικό κομμάτι. Σε μια πραγματική εφαρμογή τίθενται επιπλέον θέματα επεκτασιμότητας, παραλληλίας του αλγορίθμου καθώς και θέματα ευρωστίας. Ο αλγόριθμος στο μεγαλύτερο κομμάτι του, στα πλαίσια της ίδιας εποχής τουλάχιστον, μπορεί εύκολα να παραλληλοποιηθεί. Εργαλείο που χαρακτηρίζεται από μεγάλη άνθιση για τη μεταβίβαση των μηνυμάτων στο σύστημα επεξεργασίας είναι το kafka. Η τελική αυτή επεξεργασία μπορεί να λάβει χώρα μέσω διαφορετικών διαθέσιμων εργαλείων. Προτείνουμε τη χρήση ενός elasticsearch cluster, λόγω του γεγονότος ότι καλύπτει όλες μας τις ανάγκες. Είναι εύρωστο, γρήγορα αποκρίσιμο και δίνει τη δυνατότητα στο διαχειριστή να έχει πλήρη έλεγχο της κατανάλωσης των πόρων αυτού.

Βιβλιογραφία

[1] Aggarwal, C. C. (2017). Outlier Analysis.

[2] Babcock, B., Babu, S., Datar, M., Motwani, R., and Widom, J. (2002). Models and issues in data stream systems. In Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS '02, pages 1–16, New York, NY, USA. ACM.

[3] Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. (2005). Clustering with bregman divergences. J. Mach. Learn. Res., 6:1705–1749.

[4] Bishop, C. M. (2006). Pattern recognition and Machine Learning.

[5] Bregman, L. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Computational Mathematics and Mathematical Physics, 7(3):200–217.

[6] Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). Lof: Identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, SIGMOD '00, pages 93–104, New York, NY, USA. ACM.

[7] Cao, F., Estert, M., Qian, W., and Zhou, A. (2006). Density-based clustering over an evolving data stream with noise. In Proceedings of the 2006 SIAM International Conference on Data Mining, pages 328–339. Society for Industrial and Applied Mathematics.

[8] Cappé, O. and Moulines, E. (2009). On-line expectation-maximization algorithm for latent data models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71(3):593–613.

[9] Delimitrou, C. and Kozyrakis, C. (2014). Quasar: Resource-efficient and qos-aware cluster management. In Proceedings of the 19th International Conference

on Architectural Support for Programming Languages and Operating Systems, ASPLOS '14, pages 127–144, New York, NY, USA. ACM.

[10] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B, 39(1):1–38.

[11] Do, M. (2003). Fast approximation of kullback-leibler distance for dependence trees and hidden markov models. IEEE Signal Processing Letters, 10(4):115–118.

[12] Duda, R. O., Hart, P. E., and Stork, D. G. (2000). Pattern classification.

[13] Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996). Advances in knowledge discovery and data mining. chapter From Data Mining to Knowledge Discovery: An Overview, pages 1–34. American Association for Artificial Intelligence, Menlo Park, CA, USA.

[14] Gama, J. a., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Comput. Surv., 46(4):44:1–44:37.

[15] Ghasemi, A. and Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. International Journal of Endocrinology and Metabolism, 10(2):486–489.

[16] Hansman, S. and Hunt, R. (2005). A taxonomy of network and computer attacks. Computers & Security, 24(1):31–43.

[17] Hastie, T., Tibshirani, R., and Friedman, J. (2008). Unsupervised learning. In The Elements of Statistical Learning, pages 485–585. Springer New York.

[18] Hildebrandt, M. and Gutwirth, S. (2008). Profiling the european citizen. chapter 2, pages 17–45.

[19] Hoglund, A., Hatonen, K., and Sorvari, A. (2000). A computer host-based user anomaly detection system using the self-organizing map. In Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium. IEEE.

[20] Kejariwal, A. (2015). Introducing practical and robust anomaly detection in a time series.

[21] Kent, A. D. (2015a). Comprehensive, Multi-Source Cyber-Security Events. Los Alamos National Laboratory.

[22] Kent, A. D. (2015b). Cybersecurity Data Sources for Dynamic Network Research. In Dynamic Networks in Cybersecurity. Imperial College Press.

[23] Kim, T. K. (2015). T test as a parametric statistic. Korean Journal of Anesthesiology, 68(6):540.

[24] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. The Annals of Mathematical Statistics, 22(1):79–86.

[25] Liang, P. and Klein, D. (2009). Online em for unsupervised models. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL '09, pages 611–619, Stroudsburg, PA, USA. Association for Computational Linguistics.

[26] Liu, Z., Almhana, J., Choulakian, V., and McGorman, R. (2006). Online EM algorithm for mixture with application to internet traffic modeling. Computational Statistics & Data Analysis, 50(4):1052–1071.

[27] Manning, C. D., Raghavan, P., and Schütze, H. (2008). Introduction to Information Retrieval, chapter Flat clustering - Evaluation of clustering. Cambridge University Press, New York, NY, USA.

[28] Park, K. and Willinger, W. (2001). Self-similar Network Traffic and Performance Evaluation.

[29] Peacock, J. A. (1983). Two-dimensional goodness-of-fit testing in astronomy. Monthly Notices of the Royal Astronomical Society, 202(3):615–627.

[30] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66(336):846–850.

[31] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20:53–65.

[32] Rowland, C. H. (2002). Intrusion detection system.

[33] Schuster-Böckler, B. and Bateman, A. (2007). An introduction to hidden markov models. Appendix 3:Appendix 3A.

[34] Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2):461–464.

[35] Steyerberg, E. W., Borsboom, G. J. J. M., van Houwelingen, H. C., Eijkemans, M. J. C., and Habbema, J. D. F. (2004). Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Statistics in Medicine, 23(16):2567–2586.

[36] Syarif, I., Prugel-Bennett, A., and Wills, G. (2012). Data mining approaches for network intrusion detection: from dimensionality reduction to misuse and anomaly detection.

[37] Tan, S. C., Ting, K. M., and Liu, T. F. (2011). Fast anomaly detection for streaming data. In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two, IJCAI'11, pages 1511–1516. AAAI Press.

[38] Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2):411–423.

[39] Vlassis, N. and Likas, A. (2002). A greedy em algorithm for gaussian mixture learning. Neural Processing Letters, 15(1):77–87.

[40] Wu, K., Zhang, K., Fan, W., Edwards, A., and Yu, P. S. (2014). RS-forest: A rapid density estimator for streaming anomaly detection. In 2014 IEEE International Conference on Data Mining. IEEE.

[41] Xie, M., Hu, J., Han, S., and Chen, H.-H. (2013). Scalable hypergrid k-NN-based online anomaly detection in wireless sensor networks. IEEE Transactions on Parallel and Distributed Systems, 24(8):1661–1670.