ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΠΡΟΗΓΜΕΝΑ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Διαχείριση συμβάντων σε δεδομένα αισθητήρων πολλαπλών μεταβλητών

Αντώνης Σ. Λοΐζου

Επιβλέπων **Ευστάθιος Χατζηευθυμιάδης,** Αναπληρωτής Καθηγητής

ΑΘΗΝΑ

ΣΕΠΤΕΜΒΡΙΟΣ 2015

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Διαχείριση συμβάντων σε δεδομένα αισθητήρων πολλαπλών μεταβλητών

Αντώνης Σ. Λοΐζου
Α.Μ.: M1257

ΕΠΙΒΛΕΠΩΝ: **Ευστάθιος Χατζηευθυμιάδης,** Αναπληρωτής Καθηγητής

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Λάζαρος Μεράκος,** Καθηγητής

Σεπτέμβριος 2015

ΠΕΡΙΛΗΨΗ

Η διαχείριση συμβάντων σε δίκτυα αισθητήρων είναι ένα διεπιστημονικό πεδίο, το οποίο περιλαμβάνει στάδια σε όλη την αλυσίδα επεξεργασίας. Σε αυτή την εργασία γίνεται μία συζήτηση των βασικών βημάτων των εφαρμογών διαχείρισης συμβάντων, οι οποίες είναι πραγματικού χρόνου ή προσεγγίζουν την απόδοση εφαρμογών πραγματικού χρόνου. Τέτοιες εφαρμογές μπορεί να είναι η ανίχνευση συμβάντων, η συσχέτιση συμβάντων, η πρόβλεψη συμβάντων και το προσαρμοστικό φιλτράρισμα. Αρχικά, εξετάζονται οι υφιστάμενες εφαρμογές συστημάτων ανίχνευσης μεταβολών μίας μεταβλητής ή πολλαπλών μεταβλητών, με στόχο την ανίχνευση συμβάντων σε πραγματικό χρόνο πάνω σε δεδομένα αισθητήρων. Στη συνέχεια, γίνεται παρουσίαση ενός συστήματος συσχέτισης συμβάντων, το οποίο έχει ως στόχο την αποκάλυψη της εσωτερικής δυναμικής που διέπει τη λειτουργία ενός συστήματος. Η εσωτερική δυναμική είναι υπεύθυνη για τη δημιουργία συμβάντων, τα οποία μπορεί να διαφέρουν όσον αφορά τον τύπο τους.

Η εργασία αυτή, επίσης, παρουσιάζει την αντιπροσώπευση των εξαρτήσεων συμβάντων, η οποία μπορεί να συμπεριλαμβάνει ένα πλαίσιο αναπαράστασης γνώσης. Το πλαίσιο αυτό είναι ικανό για τη δημιουργία κανόνων συσχέτισης. Παράλληλα, στην εργασία αυτή γίνεται μία μελέτη του σημαντικού θέματος αναγνώρισης παρωχημένων εξαρτήσεων μεταξύ συμβάντων. Η αναγνώριση αυτή πραγματοποιείται μέσω της δημιουργίας ενός πλαισίου χρονικών εξαρτήσεων, το οποίο είναι ικανό να φιλτράρει τους εξαγόμενους κανόνες με την πάροδο του χρόνου. Η προτεινόμενη θεωρία εφαρμόζεται στον τομέα της ναυσιπλοΐας και επιβεβαιώνεται μέσω εκτεταμένου αριθμού πειραμάτων. Τα πειράματα αυτά πραγματοποιούνται με πραγματικές ροές αισθητήρων, οι οποίες προέρχονται από δίκτυα αισθητήρων μεγάλης κλίμακας, τα οποία είναι τοποθετημένα σε πλοία.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Διαχείριση συμβάντων

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: δίκτυα αισθητήρων, συσχέτιση συμβάντων, μαρκοβιανές αλυσίδες, ανίχνευση συμβάντων

ABSTRACT

The process of event management in sensor networks is a multidisciplinary field of science, which includes a number of steps as part of the process chain. In this thesis, we discuss the basic steps, whose execution is necessary in applications that involve management of event. These applications are characterized to be of real or near-real time. Such applications could include event detection, event correlation, event prediction and adaptive filtering. At first, existing applications of univariate or multivariate change detection systems are studied, applications that aim at detecting events in an online manner, over sensor data. Then, an event correlation engine is presented, a system that has the ability to represent the internal dynamics of a system. The internal dynamics of a system may include the creation of events, which vary in type.

This thesis also studies the representation of event dependencies, which includes a probabilistic temporal knowledge framework of representation. This framework is used for the creation of temporal association rules. In this thesis, the important topic of the recognition of old dependencies between events is also discussed. The recognition of old dependencies between events is achieved through the creation of a temporal dependencies framework, which has the ability to filter the extracted rules, based on their temporal characteristics. The proposed theory is applied in the maritime field and is evaluated through a number of experiments. These experiments are executed with real world sensor streams, which originate from big scale sensor networks deployed in ships.

SUBJECT AREA: Event Management

KEYWORDS: sensor networks, event correlation, Markov chains, event detection

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΑΛΓΟΡΙΘΜΩΝ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

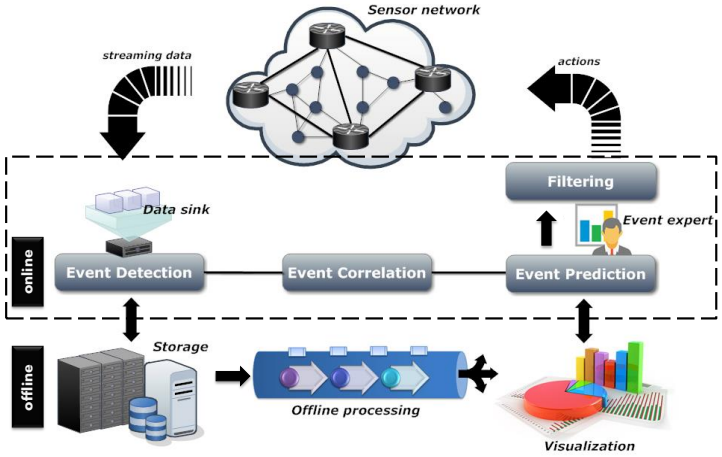
ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΓΡΑΦΙΚΩΝ ΠΑΡΑΣΤΑΣΕΩΝ

1. ΕΙΣΑΓΩΓΗ

Σε ένα δίκτυο αισθητήρων, οι κόμβοι αισθητήρων αποτελούν τα κύρια συστατικά του δικτύου, τα οποία συλλέγουν δεδομένα σχετικά με το περιβάλλον στο οποίο τοποθετούνται. Τα στοιχεία αισθητήρων συλλέγουν επίσης πληροφορία συνάφειας, η οποία χρησιμοποιείται για την υποστήριξη εφαρμογών οι οποίες λαμβάνουν υπόψη το πλαίσιο στο οποίο βρίσκονται, όπως για παράδειγμα θαλάσσιες εφαρμογές, έξυπνη μετακίνηση, έξυπνη γεωργία και άλλες. Ένα δίκτυο αισθητήρων έχει ως στόχο την ανίχνευση συμβάντων, συμβάντα τα οποία λαμβάνουν χώρα κατά τη διάρκεια ζωής του δικτύου. Τα συμβάντα αυτά μπορούν αναπαραστήσουν ή ακόμα και να επηρεάσουν την κατάσταση του συστήματος. Ο όρος συμβάν χρησιμοποιείται για την περιγραφή της μεταβολής μίας ή περισσότερων μεταβλητών, οι οποίες παρακολουθούνται από το σύστημα. Οι μεταβλητές αυτές ονομάζονται *χαρακτηριστικά πλαισίου*.

Σε δίκτυα αισθητήρων μεγάλης πολυπλοκότητας, μπορεί να γίνει διάκριση δύο κύριων ειδών επεξεργασίας όσο αφορά τα συμβάντα, η επεξεργασία συμβάντων σε πραγματικό χρόνο και η επεξεργασία συμβάντων εκτός σύνδεσης (offline). Η επεξεργασία συμβάντων σε πραγματικό χρόνο επικεντρώνεται στην ανίχνευση συμβάντων, στην αναγνώριση συσχετίσεων και αιτιατών σχέσεων με χρονικές εξαρτήσεις, στην πρόβλεψη επερχόμενων καταστάσεων του συστήματος και στο προσαρμοστικό φιλτράρισμα των δεδομένων σε πραγματικό χρόνο. Η επεξεργασία συμβάντων εκτός σύνδεσης περιλαμβάνει μεταξύ άλλων αποθήκευση συμβάντων, μεταεπεξεργασία αποθηκευμένων συμβάντων, αποθήκευση συμβάντων σε αποθήκες δεδομένων και οπτικοποίηση πληροφορίας συμβάντων. Η οπτικοποίηση πληροφορίας συμβάντων προσφέρει υποστήριξη στα συστήματα λήψης αποφάσεων για να προχωρήσουν σε ενέργειες αποκατάστασης σε περιπτώσεις που αυτό είναι απαραίτητο, όπως για παράδειγμα τροποποίηση παραμέτρων του συστήματος, αντικατάσταση ενός αισθητήρα ο οποίος είναι εκτός λειτουργίας και άλλα. Η εργασία αυτή επικεντρώνεται στην επεξεργασία συμβάντων σε πραγματικό χρόνο από δίκτυα αισθητήρων. Ακολουθείται μία σταδιακή προσέγγιση ανάλυσης των χαρακτηριστικών πλαισίου και αποκάλυψη κρυμμένων διασυνδέσεων μεταξύ των διαφόρων τύπων συμβάντων τα οποία εμφανίστηκαν κατά τη διάρκεια λειτουργίας ενός δικτύου αισθητήρων. Στην Εικόνα 1 παρουσιάζεται ο κύκλος ζωής της διαχείρισης συμβάντων σε ένα δίκτυο αισθητήρων.

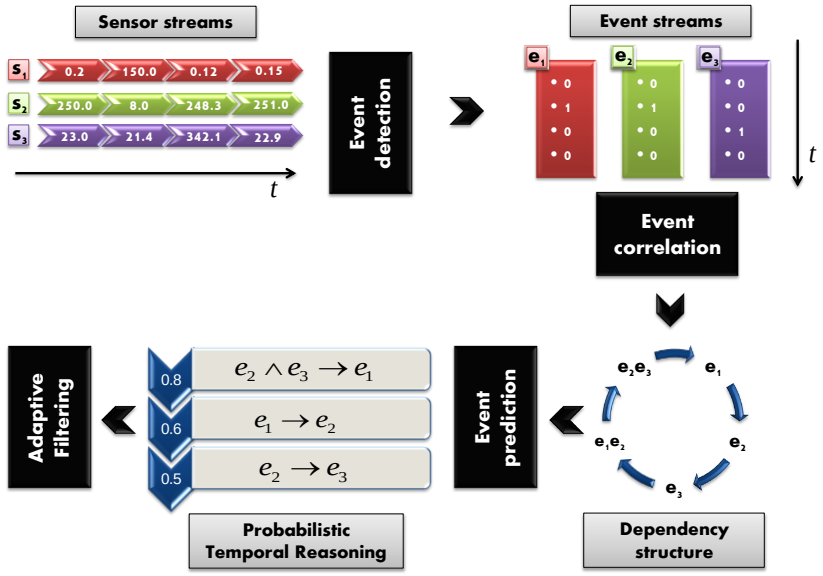


Εικόνα 1: Κύκλος ζωής διαχείρισης συμβάντων σε δίκτυα αισθητήρων

Οι ροές αισθητήρων καταφθάνουν κυρίως σε μορφή ακατέργαστων δεδομένων, τα οποία παρέχουν άμεσες μετρήσεις ή συνοψίσεις μετρήσεων όπως μέσος όρος, μέγιστην και ελάχιστη τιμή και άλλα. Οι μετρήσεις αυτές αφορούν παρατηρούμενα φαινόμενα τα οποία ενδέχεται να μεταβληθούν με την πάροδο του χρόνου. Λόγω της συχνότητάς τους, τα ακατέργαστα δεδομένα έχουν περιορισμένη αξία, ακόμα και για εμπειρογνώμονες του συστήματος, οι οποίοι έχουν ως στόχο να φτάσουν ένα υψηλότερο επίπεδο κατανόησης της εσωτερικής δυναμικής του συστήματος. Η διαχείριση και ανάλυση των ροών αισθητήρων γίνεται σε πραγματικό χρόνο. Με αυτό τον τρόπο είναι εφικτός ο προσδιορισμός χρονικών στιγμών στις οποίες η κατανομή πιθανοτήτων της χρονοσειράς μεταβάλλεται. Κάθε ροή αισθητήρων μετατρέπεται σε μία δυαδική ροή, η οποία αντιπροσωπεύει μη αναμενόμενη συμπεριφορά ενός χαρακτηριστικού πλαισίου. Η διαδικασία αυτή συμπεριλαμβάνεται στις διαδικασίες που ανήκουν στο γενικότερο πλαίσιο της συσχέτισης συμβάντων. Οι ροές συμβάντων μπορούν να παρέχουν εξευγενισμένη πληροφορία των δεδομένων αισθητήρα, η οποία μπορεί να έχει μεγάλη σημασία για τους φορείς λήψης αποφάσεων. Οι φορείς λήψης αποφάσεων πρέπει να ενεργούν με προληπτικό τρόπο, προκειμένου να διατηρηθεί η επιθυμητή συμπεριφορά του συστήματος. Για παράδειγμα, η μη φυσιολογική αύξηση των τιμών που προέρχονται από έναν ανιχνευτή καπνού θα μπορούσε ουσιαστικά να είναι ένδειξη εμφάνισης φωτιάς ή πυρκαγιάς.

Ως ένα επόμενο βήμα, η εργασία αυτή επικεντρώνεται στην παραγωγή κανόνων που περιγράφουν χρονικά εξαρτώμενους κανονισμούς μεταξύ των εισερχόμενων ροών δεδομένων αισθητήρων και στην πρόβλεψη μελλοντικών μεταβάσεων της κατάστασης του συστήματος, διαδικασία που συμπεριλαμβάνεται στο ευρύτερο πλαίσιο διαδικασιών

της πρόβλεψης συμβάντων. Οι προκύπτοντες κανόνες προέρχονται από δομές εξαρτήσεων, οι οποίες κατασκευάζονται σε προσέγγιση πραγματικού χρόνου για την αναπαράσταση αλληλεξαρτήσεων μεταξύ των μετρούμενων μεταβλητών, διαδικασία που συμπεριλαμβάνεται στο ευρύτερο πλαίσιο διαδικασιών της συσχέτισης συμβάντων. Οι προκύπτουσες εξαρτήσεις διαφορετικών χρονικών στιγμών φιλτράρονται, έτσι ώστε να γίνεται εξισορρόπηση μεταξύ παρωχημένων προτύπων και ακολουθιών συμβάντων προηγούμενων χρονικών στιγμών οι οποίες είναι ακόμα σημαντικές. Η διαδικασία αυτή συμπεριλαμβάνεται στο ευρύτερο πλαίσιο διαδικασιών του προσαρμοστικού φιλτραρίσματος. Η Εικόνα 2 παρουσιάζει τη ροή εργασίας, όπως αυτή εξετάζεται στο πλαίσιο της εργασίας αυτής, για την επεξεργασία σε πραγματικό χρόνο συμβάντων τα οποία προέρχονται από δίκτυα αισθητήρων.



Εικόνα 2: Επεξεργασία σε άμεση σύνδεση συμβάντων σε δίκτυα αισθητήρων

Η συμβολή της εργασίας αυτής είναι πολλαπλή. Αρχικά, εξετάζεται ο προσδιορισμός συμβάντων σε δίκτυα αισθητήρων με τρόπο τέτοιο έτσι ώστε να αυτά να συμπεριλαμβάνονται σε ένα πλαίσιο ανίχνευσης μεταβολών στο πλαίσιο χρονοσειρών πολλαπλών μεταβλητών. Στη συνέχεια, γίνεται περιγραφή ενός συστήματος συσχέτισης συμβάντων, το οποίο βασίζεται στην ιδέα της μερικής αντιστοίχισης (partial matching) [1], υλοποιώντας ουσιαστικά ένα μαρκοβιανό μοντέλο μεταβλητής τάξης για τη συσχέτιση των δεδομένων συμβάντων πολλαπλών μεταβλητών. Στη συνέχεια, γίνεται περιγραφή του πλαισίου για μοντελοποίηση των προκύπτοντων εξαρτήσεων συμβάντων, μέσω του προτύπου πιθανοτικού χρονικού λογικού προγραμματισμού και της θεωρίας χρονικών

κανόνων συσχέτισης. Στην εργασία αυτή γίνεται επίσης μία συζήτηση τεχνικών που ασχολούνται με χρονικές εξαρτήσεις.

Το υπόλοιπο της εργασίας αυτής οργανώνεται ως ακολούθως: στη συνέχεια της ενότητας αυτής γίνεται μία επισκόπηση των ερευνητικών προσπαθειών που σχετίζονται με τις περιοχές ανίχνευσης μεταβολών και συσχέτιση συμβάντων, όπως επίσης και μελέτη της βασικής θεωρίας δικτύων αισθητήρων. Στο Κεφάλαιο 2 γίνεται μία συζήτηση συστημάτων ανίχνευσης μεταβολών, τα οποία λαμβάνονται υπόψη στο πλαίσιο της εργασίας αυτής. Στο Κεφάλαιο 3 παρουσιάζεται μία μέθοδος συσχέτισης συμβάντων για δεδομένα συμβάντων πολλαπλών μεταβλητών και στο Κεφάλαιο 4 γίνεται συζήτηση της διαδικασίας πρόβλεψης συμβάντων μέσω της πιθανοτικής χρονικής συλλογιστικής. Στο Κεφάλαιο 5 γίνεται περιγραφή ενός πλαισίου για προσαρμοστικό φιλτράρισμα των εξαρτήσεων συμβάντων. Στο Κεφάλαιο 6 γίνεται περιγραφή του συνόλου δεδομένων, των λεπτομερειών υλοποίησης και στο Κεφάλαιο 7 παρουσιάζεται η πειραματική αξιολόγηση των προς εξέταση προσεγγίσεων. Τέλος, στο Κεφάλαιο 8 γίνεται παρουσίαση των συμπερασμάτων και των επόμενων ερευνητικών βημάτων και συνοψίζεται η διαδικασία που ακολουθήθηκε.

1.1 Προηγούμενη ερευνητική δραστηριότητα

Η διαχείριση συμβάντων έχει αποκτήσει μεγάλη δημοτικότητα τόσο στην ερευνητική περιοχή της επιστήμης της πληροφορίας, καθώς επίσης και στη βιομηχανική κοινότητα εδώ και πολλές δεκαετίες. Στην ερευνητική περιοχή της επιστήμης της πληροφορίας, τα συγγράμματα προσφέρουν ένα ευρύ φάσμα προσεγγίσεων για την ανίχνευση και το συσχετισμό συμβάντων σε συστήματα διαχείρισης δεδομένων πραγματικού χρόνου. Στο [2] οι συγγραφείς προτείνουν ένα πραγματικού χρόνου πιθανοτικό πλαίσιο για την ανίχνευση ανωμαλιών σε συνεχούς ροής αρχεία καταγραφής συμβάντων, τα οποία προέρχονται από πολλαπλές πηγές συμβάντων. Στο προτεινόμενο σύστημα, λαμβάνεται υπόψη η υπόθεση ότι τα συμβάντα τα οποία έχουν ήδη ανιχνευτεί και ο εντοπισμός των ανωμαλιών βασίζονται στην παραγωγή ενός κατευθυνόμενου πιθανοτικού γράφου. Ο γράφος αυτός αντιπροσωπεύει τη χρονική πληροφορία των εμφανίσεων συμβάντων. Οι κορυφές του γράφου, οι οποίες αντιπροσωπεύουν τους τύπους των συμβάντων, καθορίζονται από τις προηγούμενες πιθανότητες για την εμφάνιση τύπων συμβάντων, ενώ κάθε κατευθυνόμενη ακμή συνοδεύεται από μία δεσμευμένη πιθανότητα μεταξύ των συνδεόμενων τύπων συμβάντων. Οι συγγραφείς προτείνουν επίσης μία μετρική για την

ποσοτικοποίηση της ομοιότητας μεταξύ δύο ροών συμβάντων. Γίνεται ανίχνευση μίας μεταβολής όποτε αυτή η μετρική ξεπερνάει μία τιμή κατωφλίου, όταν γίνεται σύγκριση του γράφου που προέκυψε από τις πρόσφατες εγγραφές συμβάντων, δηλαδή την ουρά της ροής, με το πλήρη γράφο της αρχικής ροής.

Μία παρόμοια δομή γράφου παρουσιάζεται στο [3], όπου οι ακμές ενός αμφίδρομου γράφου αντιπροσωπεύουν τις κοινές πιθανότητες δύο τύπων συμβάντων. Οι συγγραφείς εφαρμόζουν μέτρα, τα οποία προέρχονται από τη θεωρία της πληροφορίας, προκειμένου να υπάρχει παροχή μίας ιστορικής ανάλυσης ανίχνευσης συμβάντων. Λαμβάνεται υπόψη η υπόθεση ότι τα συμβάντα είναι γνωστά και ανιχνεύονται από τα ιστορικά αρχεία καταγραφής. Οι συγγραφείς προτείνουν επίσης μετρικές και μία σταδιακή ανάλυση για τον προσδιορισμό αιτιατών συμβάντων και συμβάντων τα οποία μπορούν δυνητικά να προκαλέσουν κατάρρευση του συστήματος. Τα δύο προαναφερθέντα πλαίσια δε λαμβάνουν υπόψη ακολουθίες συσχετιζόμενων συμβάντων, αφού οι προκύπτουσες δεσμευμένες και κοινές πιθανότητες αναφέρονται σε συμβάντα τα οποία εμφανίστηκαν κατά τη διάρκεια του ίδιου χρονικού παραθύρου, αλλά όχι σε μεταγενέστερες χρονικές στιγμές. Για παράδειγμα, σε μία ροή συμβάντων όπου όποτε εμφανίζεται συμβάν τύπου Α τη χρονική στιγμή t εμφανίζεται επίσης συμβάν τύπου Β τη χρονική στιγμή $t+1$, το σύστημα δεν αποτυπώνει τη σχέση αυτή. Ομοίως, οι επιπτώσεις από την κοινή εμφάνιση συμβάντων δεν λαμβάνονται υπόψη λόγω της χρονικής πολυπλοκότητας. Για παράδειγμα, σε μία ροή συμβάντων όπου όποτε εμφανίζονται μαζί συμβάντα τύπου Α και Β τότε εμφανίζεται συμβάν τύπου C, το σύστημα δεν είναι ικανό να αποτυπώσει τη σχέση αυτή.

Στο [4] γίνεται μία παρουσίαση των κριτηρίων ανίχνευσης μεταβολών σε δεδομένα συνεχούς ροής πολλαπλών μεταβλητών. Η εργασία διερευνά ευρέως γνωστές μετρικές για ανίχνευση μεταβολών, όπως την απόσταση Kullback - Leibler [5] και τον έλεγχο T-square του Hotelling [6] για την ισοτιμία μέσων, προσφέροντας ανίχνευση μεταβολών με λογαριθμική πιθανοφάνεια. Γίνεται επίσης παρουσίαση και αξιολόγηση ενός ημιπαραμετρικού κριτηρίου λογαριθμικής πιθανοφάνειας, το οποίο φαίνεται να έχει καλύτερη απόδοση σε δεδομένα πολλαπλών μεταβλητών. Στο [7] οι συγγραφείς προτείνουν ένα πιθανοτικό σύστημα για το συσχετισμό κατανεμημένων συμβάντων στο τομέα ασφάλειας δικτύων. Συγκεκριμένα, γίνεται εφαρμογή διαδικασίας Hidden Markov Model (HMM) [8] και φιλτραρίσματος Kalman [9] για παρουσίαση χωρικών και χρονικών συσχετίσεων μεταξύ των παρατηρήσεων και των κρυμμένων καταστάσεων επιθέσεων στην ασφάλεια Διαδικτύου. Στο [10] γίνεται χρησιμοποίηση του φιλτραρίσματος Kalman

και του αλγόριθμου συσσωρευτικού αθροίσματος (CUSUM) [11] για την ανίχνευση μεταβολών σε πραγματικό χρόνο. Στην εργασία γίνεται επίσης αξιολόγηση διαφόρων κριτηρίων για την ανίχνευση μεταβολών σε δεδομένα συνεχούς ροής.

Ο τομέας της βιομηχανικής συντήρησης έχει επίσης αποκτήσει ιδιαίτερο ενδιαφέρον στην ακαδημαϊκή βιβλιογραφία για πολλές δεκαετίες. Συγκεκριμένα, πολλές μελέτες έχουν επικεντρωθεί στην περιοχή της προβλεπτικής συντήρησης [12], η οποία βασίζεται σε καταστάσεις για την πρόβλεψη βλαβών και σφαλμάτων σε πολύπλοκα τεχνικά συστήματα, συμπεριλαμβανομένων και των δικτύων αισθητήρων. Οι περισσότερες των βιομηχανικών προσεγγίσεων βασίζονται σε υπάρχουσες στατιστικές μεθόδους για την ανίχνευση καταστάσεων συναγερμού, οι οποίες οφείλονται σε απότομες μεταβολές των κατανομών πιθανοτήτων των παρατηρούμενων παραμέτρων [13].

Στο [14] γίνεται συζήτηση του προβλήματος πρόβλεψης σπάνιων συμβάντων σε δεδομένα πολλαπλών μεταβλητών. Επίσης προτείνονται δύο προσεγγίσεις για την αντιμετώπιση του προβλήματος. Η πρώτη προσέγγιση επικεντρώνεται στην ανίχνευση υποβάθμισης, δηλαδή της μη φυσιολογικής συμπεριφοράς του συστήματος. Η ανίχνευση αυτή βασίζεται στο μοντέλο Μηχανών Υποστήριξης Διανυσμάτων μίας κλάσης [15]. Συγκεκριμένα, γίνεται ανάπτυξη ενός προγράμματος, το οποίο κάνει ανίχνευση ανωμαλιών με την ελαχιστοποίηση της απόστασης κάθε νέου πολυδιάστατου διανύσματος δεδομένων σε σχέση με ένα σύνολο δεδομένων εκπαίδευσης. Στη συνέχεια, χρησιμοποιείται ένα μοντέλο κινητού μέσου όρου για την μοντελοποίηση πολυδιάστατης συμπεριφοράς αποικοδόμησης στο χρόνο. Η δεύτερη προσέγγιση εκμεταλλεύεται μία μέθοδο κατηγοριοποίησης κανονικοποιημένης λογιστικής παλινδρόμησης, στην οποία η συνάρτηση πρόβλεψης αποτελείται από ένα μετασχηματισμένο γραμμικό συνδυασμό των επεξηγηματικών μεταβλητών. Και οι δύο προσεγγίσεις αξιολογούνται πάνω σε σύνολα δεδομένων τα οποία προέκυψαν από περιπτώσεις πραγματικού κόσμου από την περιοχή απόδοσης επιχειρήσεων αεροσκαφών.

Στο [16] οι συγγραφείς προτείνουν μία μέθοδο βασισμένη στο μοντέλο και μία μέθοδο βασισμένη στα δεδομένα για την πρόβλεψη βλαβών σε συστήματα πολλαπλών αισθητήρων. Η πρώτη προσέγγιση βασίζεται σε ένα αυτοπαλίνδρομο μοντέλο κινητού μέσου όρου, όπου οι εκτιμώμενες παράμετροι του μοντέλου χρησιμοποιούνται για την υλοποίηση του ανιχνευτή μεταβολών. Ο ανιχνευτής μεταβολών υλοποιείται ως ένας έλεγχος υποθέσεων Neyman-Pearson [17]. Η δεύτερη προσέγγιση χρησιμοποιεί τους αλγόριθμους HMM και Viterbi [18] για την εκτίμηση της πιο πιθανής ακολουθίας

κρυμμένων καταστάσεων. Και οι δύο προσεγγίσεις αξιολογούνται μέσω προσομοιώσεων με συνθετικά δεδομένα.

1.2 Σήμα

Ένα σήμα είναι μία σειρά από αριθμούς, οι οποίοι προκύπτουν από ένα μέγεθος [19]. Οι αριθμοί αυτοί τυπικά λαμβάνονται χρησιμοποιώντας κάποια μέθοδο καταγραφής σε συνάρτηση με το χρόνο. Σε πραγματικές εφαρμογές, τα σήματα μπορούν να κατηγοριοποιηθούν σε δύο ομάδες, τα σήματα με στάσιμη συμπεριφορά και τα σήματα με μη στάσιμη συμπεριφορά. Ένα σήμα το οποίο αντιπροσωπεύει ένα τυχαίο φαινόμενο μπορεί να χαρακτηριστεί με συμπεριφορά είτε στατική είτε μη στατική. Οι στατιστικές ιδιότητες των στάσιμων σημάτων παραμένουν αμετάβλητες στο χρόνο. Αντίθετα, οι στατιστικές ιδιότητες των μη στάσιμων σημάτων εξαρτώνται από τη χρονική στιγμή της καταμέτρησης.

1.3 Απότομες μεταβολές

Οι απότομες μεταβολές ορίζονται ως οι ραγδαίες αλλαγές, που προκύπτουν σε σχέση με την περίοδο δειγματοληψίας των μετρήσεων [13]. Η διαδικασία ανίχνευσης απότομων αλλαγών συμπεριλαμβάνει εργαλεία, τα οποία βοηθούν στη λήψη απόφασης μεταβολής στα χαρακτηριστικά του εξεταζόμενου μεγέθους. Το γενικό πρόβλημα ανίχνευσης απότομων μεταβολών στις παραμέτρους διαδικασιών έχει ευρέως μελετηθεί στη βιβλιογραφία. Οι μεταβολές αυτές μπορεί να οφείλονται σε μετατοπίσεις που μπορεί να εμφανιστούν στη μέση τιμή ή στη μεταβολή της δυναμικής του σήματος.

1.4 Δίκτυα αισθητήρων

Τα δίκτυα αισθητήρων [20] [21] μπορεί να αποτελούνται από διαφορετικών τύπων αισθητήρες, όπως θερμικοί αισθητήρες, οπτικοί αισθητήρες, αισθητήρες υπερύθρων και άλλα. Η ποικιλομορφία τύπων αισθητήρων σε ένα δίκτυο αισθητήρων επιτρέπει την παρακολούθηση ενός ευρέος φάσματος συνθηκών του περιβάλλοντος, όπως η θερμοκρασία, η υγρασία, η ατμοσφαιρική πίεση, ο θόρυβος στο περιβάλλον, η ταχύτητα του ανέμου και άλλα. Τα δίκτυα αισθητήρων μπορούν να ταξινομηθούν σε δύο κατηγορίες όσον αφορά την εφαρμογή τους: την παρακολούθηση (monitoring) και τον εντοπισμό

(detection). Εφαρμογές παρακολούθησης μπορεί να είναι: ανίχνευση κινήσεων εχθρού σε στρατιωτικές επιχειρήσεις, παρακολούθηση διαθέσιμων αποθεμάτων σε επιχειρήσεις, παρακολούθηση κινήσεων ζώων, παρακολούθηση ασθενών σε νοσοκομεία και άλλα. Εφαρμογές ανίχνευσης μπορεί να είναι: ανίχνευση εχθρού σε στρατιωτικές επιχειρήσεις, ανίχνευση ζώων, ανίχνευση κίνησης σε αυτοκινητόδρομους, ανίχνευση θέσης αυτοκινήτου ή λεωφορείου και άλλα.

Ένας κόμβος αισθητήρα είναι ένας κόμβος σε ένα δίκτυο αισθητήρων ο οποίος είναι ικανός να εκτελεί κάποιου είδους επεξεργασία, να συγκεντρώνει αισθητηριακή πληροφορία και να επικοινωνεί με άλλους κόμβους, οι οποίοι είναι συνδεδεμένοι στο δίκτυο. Κάθε κόμβος στο δίκτυο αισθητήρων αποτελείται από τέσσερις κύριες μονάδες, την αισθητήρια μονάδα, τη μονάδα επεξεργασίας δεδομένων, τη μονάδα ενέργειας και τη μονάδα μετάδοσης δεδομένων [20].

1.4.1 Ανίχνευση ανωμαλιών και συμβάντων σε δίκτυα αισθητήρων

Η λήψη αποφάσεων, η οποία να είναι ορθή με βάση τα δεδομένα από κόμβους αισθητήρων, σε ένα δίκτυο αισθητήρων προϋποθέτει την ποιότητα των δεδομένων από τους κόμβους αισθητήρων. Τα μοντέλα ανίχνευσης ανωμαλιών σε δίκτυα αισθητήρων μπορούν να διασφαλίσουν την ποιότητα των δεδομένων αυτών από κινδύνους, όπως οι επιθέσεις εκ των έσω για παραποίηση των δεδομένων. Τα μοντέλα ανίχνευσης ανωμαλιών σχεδιάζονται για ανίχνευση μη φυσιολογικής συμπεριφοράς σε ροές δεδομένων αισθητήρων [22].

Σε γενικό επίπεδο, ανωμαλίες ορίζονται ως «πρότυπα στα δεδομένα, τα οποία δεν είναι σύμφωνα με μία καλά καθορισμένη έννοια της φυσιολογικής συμπεριφοράς» [23]. Στα δίκτυα αισθητήρων, οι ανωμαλίες μπορούν να οριστούν ως οι σημαντικές αποκλίσεις στα δεδομένα αισθητήρων των μετρήσεων από ένα προφίλ, το οποίο περιγράφει φυσιολογικά δεδομένα αισθητήρων [22] [24]. Οι ανωμαλίες σε δίκτυα αισθητήρων μπορούν να εμφανιστούν για διάφορους λόγους, όπως σφάλματα στις μετρήσεις λόγω ελαττωματικών κόμβων αισθητήρων, θόρυβος που προκύπτει από εξωτερικούς παράγοντες, συμβάντα τα οποία προκύπτουν από το περιβάλλον, κακόβουλες επιθέσεις μέσω παραβιασμένων κόμβων αισθητήρων και άλλα. Η ανίχνευση ανωμαλιών ορίζεται ως το πρόβλημα εύρεσης, συνήθως σε πραγματικό χρόνο, προτύπων στα δεδομένα, τα

οποία δεν ανήκουν σε μία αναμενόμενη συμπεριφορά, όπως αυτή προκύπτει από τη μελέτη των δεδομένων.

Η χρήση των δικτύων αισθητήρων για παρακολούθηση φαινομένων, όπως είναι για παράδειγμα καιρικά φαινόμενα ή ανίχνευση πυρκαγιών, είναι ένα κίνητρο για εφαρμογή ανίχνευσης ανωμαλιών στα δεδομένα αισθητήρων που προκύπτουν, και συγκεκριμένα εφαρμογή ανίχνευσης συμβάντων. Η εφαρμογή ανίχνευσης ανωμαλιών, και συγκεκριμένα ανίχνευσης συμβάντων για παρακολούθηση περιβάλλοντος και φαινομένων μπορεί να βοηθήσει στην ανίχνευση κάποιου σοβαρού προβλήματος ή κάποιας σοβαρής καταστροφής, όπως για παράδειγμα πυρκαγιά, στα πρώιμα στάδια. Επίσης, η ανίχνευση συμβάντων μπορεί να βοηθήσει στη λήψη αποφάσεων στις περιπτώσεις αυτές. Τέλος, κάποιο συμβάν μπορεί να παρουσιάζει ενδιαφέρον για περαιτέρω ανάλυση και μελέτη.

1.4.2 Χαρακτηριστικά δεδομένων αισθητήρων

Τα δεδομένα αισθητήρων συλλέγονται με τη μορφή ροών δεδομένων, οι οποίες προκύπτουν από πραγματικές παρατηρήσεις που συλλέγονται από το περιβάλλον. Κάποια δίκτυα αισθητήρων σχεδιάζονται για συλλογή ενός είδους δεδομένων, ενώ κάποια άλλα δίκτυα αισθητήρων σχεδιάζονται για συλλογή πολλών ειδών δεδομένων από το περιβάλλον ταυτοχρόνως, όπως για παράδειγμα θερμοκρασία, υγρασία, πληροφορία ανέμου και άλλα. Στην πρώτη περίπτωση, τα δεδομένα αισθητήρων που προκύπτουν ονομάζονται δεδομένα μιας μεταβλητής, ενώ στη δεύτερη περίπτωση τα δεδομένα αισθητήρων που προκύπτουν ονομάζονται δεδομένα πολλαπλών μεταβλητών.

Για συλλογή δεδομένων αισθητήρων πολλαπλών μεταβλητών, οι κόμβοι στις περισσότερες των περιπτώσεων περιέχουν περισσότερες από μία αισθητήριες μονάδες για ταυτόχρονη συλλογή δεδομένων διαφορετικών τύπων. Κάθε είδος δεδομένων στα δεδομένα πολλαπλών μεταβλητών ονομάζεται χαρακτηριστικό. Η ανίχνευση συμβάντων στα δεδομένα μίας μεταβλητής επιτυγχάνεται με παρακολούθηση κάποιας τιμής του χαρακτηριστικού σε σύγκριση με τις άλλες τιμές του χαρακτηριστικού σε διαδοχικές ή κοντινές χρονικές στιγμές.

Η ανίχνευση συμβάντων στα δεδομένα πολλαπλών μεταβλητών μπορεί να επιτευχθεί με παρόμοιο τρόπο για κάθε χαρακτηριστικό ξεχωριστά, με τις μεθόδους ανίχνευσης συμβάντων μιας μεταβλητής, αλλά και με τρόπο που λαμβάνει υπόψη ένα

υποσύνολο χαρακτηριστικών, με τις μεθόδους ανίχνευσης συμβάντων πολλαπλών μεταβλητών. Οι μέθοδοι ανίχνευσης συμβάντων πολλαπλών μεταβλητών είναι χρήσιμες γιατί σε δεδομένα πολλαπλών μεταβλητών, οι τιμές κάθε μεταβλητής μπορεί να μην παρουσιάζουν κάποια μη φυσιολογική συμπεριφορά, αλλά όταν λαμβάνονται υπόψη μαζί είναι πιθανό να παρουσιάζουν μη φυσιολογική συμπεριφορά.

Παραδείγματα μεθόδων ανίχνευσης συμβάντων μίας μεταβλητής είναι ο αλγόριθμος διαγραμμάτων συσσωρευτικού αθροίσματος (CUSUM) [11] και ο αλγόριθμος διαγραμμάτων ελέγχου Shewhart [25], ενώ παραδείγματα μεθόδων ανίχνευσης συμβάντων πολλαπλών μεταβλητών είναι η παλινδρόμηση πολλαπλών μεταβλητών [26] και η στατιστική σάρωση Bayes πολλαπλών μεταβλητών [27]. Παρόλο που η ανίχνευση συμβάντων σε δεδομένα πολλαπλών μεταβλητών είναι υπολογιστικά δαπανηρή, η ανίχνευση συμβάντων σε δεδομένα πολλαπλών μεταβλητών μπορεί να έχει υψηλή ακρίβεια, εάν γίνει αποδοτική εκμετάλλευση των συσχετίσεων μεταξύ των επιμέρους χαρακτηριστικών.

Τα δεδομένα αισθητήρων μπορούν να εμφανίζουν χωρικές και χρονικές εξαρτήσεις μεταξύ των μετρήσεων. Η χωρική συσχέτιση ερμηνεύεται ως η υπόθεση ότι οι μετρήσεις των κόμβων, οι οποίοι βρίσκονται σε γεωγραφικά κοντινή απόσταση μεταξύ τους, συσχετίζονται. Η χρονική συσχέτιση ερμηνεύεται ως η υπόθεση ότι οι μετρήσεις, οι οποίες προέκυψαν μία χρονική στιγμή, συσχετίζονται με τις μετρήσεις οι οποίες προέκυψαν σε μία άλλη παρελθοντική χρονική στιγμή.

1.4.3 Ανίχνευση συμβάντων σε δίκτυα αισθητήρων

Σε ένα δίκτυο αισθητήρων, ο εντοπισμός συμπεριφορών οι οποίες δε συνάδουν με το «φυσιολογικό», όπου φυσιολογικό θα επεξηγηθεί στο περαιτέρω κείμενο, είναι μία σημαντική πρόκληση για εφαρμογές όπως η παρακολούθηση, η διάγνωση βλαβών και η ανίχνευση εισβολής εξωτερικού παράγοντα. Μία σημαντική πρόκληση είναι η ελαχιστοποίηση της επικοινωνίας και της κατανάλωσης ενέργειας κατά τη διαδικασία εύρεσης μη φυσιολογικών συμπεριφορών στο δίκτυο. Τα δίκτυα αισθητήρων αποτελούνται από ένα μεγάλο αριθμό κόμβων μικρού μεγέθους, οι οποίοι χαρακτηρίζονται από περιορισμένους πόρους ενέργειας, εύρους ζώνης, μνήμης και υπολογιστικής ισχύος. Οι εγγενείς αυτοί περιορισμοί των κόμβων αισθητήρων μπορούν να κάνουν ένα δίκτυο πιο ευάλωτο σε σφάλματα και κακόβουλες επιθέσεις. Η βασική

πρόκληση για τον εντοπισμό μη φυσιολογικών συμπεριφορών σε δίκτυα αισθητήρων είναι η ανάπτυξη αλγορίθμων οι οποίοι να είναι ικανοί να ανιχνεύουν ανωμαλίες στο δίκτυο [28]. Η ανίχνευση αυτή πρέπει γίνεται με τέτοιο τρόπο ώστε η επικοινωνία μεταξύ των κόμβων και η κατανάλωση ενέργειας στο δίκτυο να ελαχιστοποιούνται.

Μη φυσιολογικές συμπεριφορές σε ένα δίκτυο μπορούν να αναγνωριστούν με την ανάλυση των δεδομένων από τους κόμβους αισθητήρων. Ένας κόμβος μπορεί να εμφανίσει μη φυσιολογική συμπεριφορά όταν προκύπτει κάποιο σφάλμα, όταν συμβαίνει κάτι στο περιβάλλον το οποίο δε συνάδει με τις φυσιολογικές συνθήκες, ή όταν κάποιοι αισθητήρες παραβιαστούν, με αποτέλεσμα να λειτουργούν κακόβουλα. Σε κάθε μία από τις παραπάνω περιπτώσεις, οι μη φυσιολογικές συμπεριφορές μπορούν να εντοπιστούν με την ανάλυση των μετρήσεων των αισθητήρων, με σκοπό τη διάκριση της κανονικής από τη μη κανονική συμπεριφορά. Η υποκείμενη διασύνδεση μεταξύ των μετρήσεων των αισθητήρων στις περισσότερες περιπτώσεις δεν είναι γνωστή εκ των προτέρων. Συνεπώς, η ανίχνευση συμβάντων σε δεδομένα με άγνωστην κατανομή είναι ένα σημαντικό πρόβλημα στα δίκτυα αισθητήρων [21].

Ένα συμβάν ορίζεται ως μία παρατήρηση η οποία φαίνεται να είναι ασυμβίβαστη με τις υπόλοιπες τιμές ενός συνόλου δεδομένων μίας μέτρησης. Η ανίχνευση συμβάντων ορίζεται ως η διαδικασία εύρεσης προτύπων στα δεδομένα, τα οποία αποκλίνουν από την αναμενόμενη συμπεριφορά. Οι διάφορες προσεγγίσεις ανίχνευσης συμβάντων στα δίκτυα αισθητήρων χαρακτηρίζονται από την αποτελεσματικότητα με την οποία κάνουν ανίχνευση, όπως επίσης και από την αποδοτικότητα με την οποία χρησιμοποιούν τους περιορισμένους πόρους του δικτύου. Η αποτελεσματικότητα ανίχνευσης αποτελείται από ένα σύνολο από μετρικές, όπως η ακρίβεια ανίχνευσης, το ποσοστό ανίχνευσης και ο αριθμός των λανθασμένων συναγερμών. Η αποδοτικότητα ανίχνευσης αποτελείται από μετρικές που περιγράφουν την κατανάλωση ενέργειας και μνήμης. Μία ιδανική προσέγγιση ανίχνευσης συμβάντων σε ένα σύστημα ασύρματων αισθητήρων πρέπει να μεγιστοποιεί την αποτελεσματικότητα ανίχνευσης συμβάντων και να ελαχιστοποιεί την κατανάλωση ενέργειας και αποθηκευτικών πόρων κατά τη διαδικασία της ανίχνευσης συμβάντων.

1.4.4 Προκλήσεις στην ανακάλυψη γνώσης σε δίκτυα αισθητήρων

Μία ροή δεδομένων είναι μια διατεταγμένη ακολουθία στοιχείων, η οποία λαμβάνεται με κάποια χρονική συμπεριφορά. Σε αντίθεση με τα δεδομένα που λαμβάνονται από στατικές Βάσεις Δεδομένων, οι ροές δεδομένων είναι συνεχείς, απεριόριστες, συνήθως λαμβάνονται με υψηλές ταχύτητες και χαρακτηρίζονται από μεταβαλλόμενη με το χρόνο κατανομή δεδομένων.

Λόγω των χαρακτηριστικών των δεδομένων ροής, υπάρχουν κάποιες εγγενείς προκλήσεις στην ανακάλυψη γνώσης από δεδομένα αισθητήρων [29] [30]. Λόγω των χαρακτηριστικών των ροών δεδομένων, συνεχή, απεριόριστα και με υψηλή ταχύτητα, υπάρχει μία μεγάλη ποσότητα δεδομένων στις ροές δεδομένων, οπότε η διαδικασία επανασάρωσης δεν είναι πρακτικά εφικτή όταν συμβαίνει μία ενημέρωση. Επίσης, δεν υπάρχει αρκετός χώρος για αποθήκευση όλων των δεδομένων αισθητήρων για επεξεργασία σε πραγματικό χρόνο. Συνεπώς, στην ανακάλυψη γνώσης από δεδομένα αισθητήρων είναι απαραίτητη η μία σάρωση των δεδομένων και η σωστή και συμπαγής χρήση της μνήμης. Επίσης, είναι απαραίτητη η σωστή προσαρμογή στη μεταβαλλόμενη κατανομή δεδομένων, σε διαφορετική περίπτωση υπάρχει το ενδεχόμενο να εμφανιστεί το πρόβλημα μετατόπισης εννοιών. Στην περίπτωση των δεδομένων αισθητήρων σε πραγματικό χρόνο, υπάρχει η ανάγκη επεξεργασίας τους σε σύντομο χρονικό διάστημα. Συγκεκριμένα, η ταχύτητα της διαδικασίας ανακάλυψης γνώσης πρέπει να είναι μεγαλύτερη από την ταχύτητα άφιξης δεδομένων, σε διαφορετική περίπτωση πρέπει να γίνει εφαρμογή μεθόδων προσέγγισης δεδομένων, όπως δειγματοληψία και απόρριψη φορτίου, μέθοδοι οι οποίοι οδηγούν σε μείωση της ακρίβειας αποτελεσμάτων. Ένα ακόμα χαρακτηριστικό πρέπει να είναι ο αυξητικός χαρακτήρας των αποτελεσμάτων, με άλλα λόγια τα αποτελέσματα πρέπει να βασίζονται σε αποτελέσματα προηγούμενων χρονικών στιγμών. Επίσης, πρέπει να γίνεται προσαρμογή της μεθόδου με τους διαθέσιμους πόρους μνήμης και υπολογιστικής δύναμης.

1.4.5 Ανοικτά θέματα στην ανίχνευση συμβάντων σε δίκτυα αισθητήρων

Τα μοντέλα ανίχνευσης συμβάντων σε δίκτυα αισθητήρων πρέπει να χρησιμοποιούν τεχνικές όπως η μείωση διαστάσεων, η σε πραγματικό χρόνο ανίχνευση, η κατανεμημένη ανίχνευση, η προσαρμοστική ανίχνευση και η εκμετάλλευση της συσχέτισης μεταξύ των δεδομένων [21] [31]. Είναι σημαντική η απαίτηση ένα μοντέλο ανίχνευσης συμβάντων σε δίκτυα αισθητήρων να κάνει ανίχνευση σε πραγματικό χρόνο, αλλά και με κατανεμημένο

τρόπο στο δίκτυο. Η σε πραγματικό χρόνο ανίχνευση διασφαλίζει ότι δεν υπάρχει απώλεια ανωμαλιών, οι οποίες εμφανίζονται σε πραγματικό χρόνο, ενώ η ανίχνευση με κατανεμημένο χαρακτήρα διασφαλίζει ότι οι περιορισμένοι πόροι του δικτύου αισθητήρων χρησιμοποιούνται με αποδοτικό τρόπο. Ο αποδοτικός τρόπος στην περίπτωση αυτή ερμηνεύεται ως η κατανομή του υπολογιστικού φόρτου στους κόμβους του δικτύου, με άλλα λόγια η αποφυγή κεντρικοποιημένης επεξεργασίας σε κάποιους κόμβους του δικτύου. Η μείωση διαστάσεων στοχεύει στην μείωση των διαστάσεων των δεδομένων με σκοπό την αύξηση της αποδοτικότητας του συστήματος. Η προσαρμοστική ανίχνευση είναι απαραίτητη για ανίχνευση σε πραγματικό χρόνο σε περιβάλλοντα τα οποία έχουν δυναμικό χαρακτήρα. Η συσχέτιση των δεδομένων χρησιμοποιώντας κάποια τεχνική κατηγοριοποίησης, όπως κάποια παραλλαγή του αλγορίθμου κοντινότερων γειτόνων, στοχεύει στην αύξηση της αποτελεσματικότητα ανίχνευσης του συστήματος. Στη συνέχεια γίνεται μια μελέτη των τεχνικών αυτών.

1.4.5.1 Μείωση Διαστάσεων

Η μείωση διαστάσεων είναι χρήσιμη στα δεδομένα πολλαπλών μεταβλητών γιατί μπορεί να οδηγήσει σε μείωση κατανάλωσης ενέργειας στους επιμέρους κόμβους του δικτύου. Λόγω της υπολογιστικής πολυπλοκότητας κάποιων μεθόδων μείωσης διαστάσεων, η οποία μπορεί τελικά να οδηγήσει σε αύξηση κατανάλωσης ενέργειας, απαιτούνται μέθοδοι οι οποίοι να είναι υπολογιστικά λιγότερο κοστοβόροι.

1.4.5.2 Ανίχνευση σε πραγματικό χρόνο

Είναι προτιμότερο η ανίχνευση συμβάντων να είναι σε πραγματικό χρόνο, έτσι ώστε να ικανοποιούνται οι απαιτήσεις ορισμένων εφαρμογών δικτύων αισθητήρων, οι οποίες λειτουργούν σε πραγματικό χρόνο. Κάποιες μέθοδοι ανίχνευσης σε πραγματικό χρόνο μπορούν να οδηγήσουν σε αύξηση κατανάλωσης ενέργειας. Για το λόγο αυτό, απαιτούνται μέθοδοι ανίχνευσης σε πραγματικό χρόνο οι οποίες να είναι υπολογιστικά λιγότερο κοστοβόρες.

1.4.5.3 Κατανεμημένη δομή

Η ανίχνευση συμβάντων είναι προτιμότερο να γίνεται με κατανεμημένο τρόπο, έτσι ώστε να αποφεύγεται η υπολογιστική φόρτωση κάποιων κόμβων, δηλαδή άνισος διαμοιρασμός υπολογιστικού φόρτου μέσα στο δίκτυο. Επίσης με την κατανεμημένη δομή μειώνεται το κόστος επικοινωνίας μεταξύ των κόμβων του δικτύου, όταν αυτοί πρέπει να αποστέλλουν τα δεδομένα σε κεντρικοποιημένες περιοχές για επεξεργασία.

1.4.5.4 Προσαρμοστική ανίχνευση

Λόγω του δυναμικού χαρακτήρα των δεδομένων αισθητήρων των κόμβων, το μοντέλο ανίχνευσης συμβάντων πρέπει να προσαρμόζεται στις πιθανές αλλαγές. Η προσαρμοστική ανίχνευση μπορεί να οδηγήσει σε μείωση του αριθμού των λανθασμένων συναγερμών και συνεπώς αύξηση της ακρίβειας πρόβλεψης. Οι μέθοδοι προσαρμοστικής ανίχνευσης είναι στις περισσότερες περιπτώσεις υπολογιστικά κοστοβόρες, ως εκ τούτου απαιτείται κάποια μορφοποίηση τους, έτσι ώστε να είναι λιγότερο κοστοβόρες και συνεπώς η κατανάλωση ενέργειας να μην είναι μεγάλη.

1.4.5.5 Συσχετίσεις χαρακτηριστικών

Η εκμετάλλευση χωρικών και χρονικών συσχετίσεων των δεδομένων αισθητήρων μπορεί να οδηγήσει σε αύξηση τόσο της αποτελεσματικότητας όσο και της αποδοτικότητας ανίχνευσης συμβάντων. Επίσης, η εκμετάλλευση των συσχετίσεων μεταξύ των χαρακτηριστικών, στην περίπτωση των δεδομένων αισθητήρων πολλαπλών μεταβλητών, μπορεί να οδηγήσει σε παρόμοια αποτελέσματα. Για το λόγο αυτό, η εκμετάλλευση χωρικών και χρονικών συσχετίσεων, όπως επίσης και συσχετίσεων χαρακτηριστικών, θεωρείται ένα σημαντικό χαρακτηριστικό το οποίο πρέπει να υπάρχει σε μία μέθοδο ανίχνευσης συμβάντων.

1.4.5.6 Ρύθμιση παραμέτρων

Οι παράμετροι, οι οποίοι αποτελούν είσοδο σε μία μέθοδο ανίχνευσης συμβάντων, αποτελούν ένα σημαντικό παράγοντα επηρεασμού της αποτελεσματικότητας και αποδοτικότητας του μοντέλου. Τέτοιοι παράμετροι μπορεί να είναι ο μέσος όρος, η τυπική απόκλιση, η ολίσθηση, οι τιμές κατωφλίου, οι τιμές συσχετίσεων μεταξύ

μετρήσεων και άλλες. Είναι επιθυμητό οι παράμετροι αυτοί να προσεγγίζονται από το ίδιο το μοντέλο, έτσι ώστε να αποφεύγεται η χειροκίνητη αρχικοποίηση τους, η οποία μπορεί να οδηγήσει σε μείωση της αποτελεσματικότητας.

2. ΑΝΙΧΝΕΥΣΗ ΣΥΜΒΑΝΤΩΝ ΣΕ ΡΟΕΣ ΔΕΔΟΜΕΝΩΝ ΑΙΣΘΗΤΗΡΩΝ ΠΟΛΛΑΠΛΩΝ ΜΕΤΑΒΛΗΤΩΝ

Σε αυτή την ενότητα μελετάται η διαδικασία ανίχνευσης συμβάντων και παραγωγής ροών συμβάντων πάνω σε ένα υπάρχον σύνολο ροών αισθητήρων. Το πρόβλημα ανίχνευσης συμβάντων πάνω σε πολλαπλές ροές αισθητήρων μπορεί να διαμορφωθεί όπως περιγράφεται στη συνέχεια. Αρχικά γίνεται παρατήρηση σε πραγματικό χρόνο με κάποια ενιαία συχνότητα χρονοσειρών πολλαπλών μεταβλητών των ποσοτικών παραμέτρων απόδοσης του συστήματος. Μία ροή αισθητήρων, η οποία αποτελείται από αριθμητικές τιμές αισθητήρων συμβολίζεται με s_i και με $s_i(t)$ συμβολίζεται η τιμή της ροής s_i σε χρόνο t, όπου ισχύει $t\epsilon[0,+\infty)$. Υποθέτοντας ότι n ροές αισθητήρων συγχρονίζονται για την αναφορά των τιμών τους περιοδικά, γίνεται αντιπροσώπευση του συνόλου πληροφορίας πλαισίου πολλαπλών μεταβλητών σε κάθε χρονική στιγμή t με ένα διάνυσμα πλαισίου $\Delta\Pi_t = (s_1(t), s_2(t), \dots, s_n(t))\epsilon\Re^n$. Πρακτικά, κάθε ροή αισθητήρων διαμορφώνει μία μονοδιάστατη χρονοσειρά, ενώ η ροή διανυσμάτων πλαισίου αντιπροσωπεύει μία χρονοσειρά πολλαπλών μεταβλητών.

Υπάρχουν πολλά προβλήματα σε τομείς επιστημών, τα οποία απαιτούν την ακολουθιακή ανίχνευση μίας αλλαγής ή ενός συμβάντος σε μία διαδικασία. Στην πιο απλή μορφή, γίνεται προσπάθεια για ανίχνευση μίας αλλαγής στο μέσο όρο μίας ακολουθίας, όπου η αλλαγή είναι είτε απότομη ή σταδιακή. Μία ροή δεδομένων αποτελείται από μία δυνητικά άπειρη ακολουθία πλειάδων δεδομένων. Τα δεδομένα ροής έχουν δύο χαρακτηριστικά, τα οποία αποτελούν πρόκληση στην επεξεργασία τους, ο υψηλός ρυθμός άφιξης και το ενδεχόμενο μη προβλέψιμης συμπεριφοράς.

Η ανίχνευση συμβάντων πάνω σε ροές αισθητήρων έχει ως στόχο το προσδιορισμό των τιμών $s_i(t)$, οι οποίες αποτελούν απότομες μεταβολές μέσα σε μία ροή διανυσμάτων πλαισίου. Συγκεκριμένα, κάθε διάνυσμα πλαισίου μήκους n μετατρέπεται σε ένα δυαδικό διάνυσμα του ιδίου μήκους, με κάθε τιμή να αντιπροσωπεύει κάποια πιθανή μεταβολή στην αντίστοιχη ροή αισθητήρων. Τέτοιες αποκλίσεις από τη φυσιολογική συμπεριφορά ονομάζονται συμβάντα και τα δυαδικά διανύσματα ονομάζονται διανύσματα συμβάντων.

Ένα συμβάν μπορεί να είναι μία παρατήρηση η οποία δεν είναι σύμφωνη με ένα αναμενόμενο πρότυπο στο σύνολο δεδομένων. Τα συμβάντα μπορεί να έχουν προκληθεί από διάφορους λόγους, όπως για παράδειγμα βλάβη ή δυσλειτουργία στους αισθητήρες, τιμές απόκλισης ή ουσιαστικές αλλαγές οι οποίες μπορεί να επηρεάσουν τη συμπεριφορά του συστήματος. Ως εκ τούτου, ένα διάνυσμα συμβάντων σε χρόνο t αντιπροσωπεύεται

από $\Delta\Sigma_t = (e_1^t, e_2^t, \dots, e_n^t)\epsilon\{0,1\}^n$ όπου $e_i^t = e_i(t)$ είναι η δυαδική τιμή η οποία αντιπροσωπεύει κατά πόσο εμφανίστηκε μια μη φυσιολογική συμπεριφορά στη ροή, η οποία αντιπροσωπεύεται με τιμή ίση με ένα, σε χρόνο t ή τιμή $s_i(t)$ συμπεριλαμβανόταν στο αναμενόμενο εύρος τιμών.

Η μετατροπή ενός διανύσματος πλαισίου σε ένα διάνυσμα συμβάντων βασίζεται σε αλγόριθμους ανίχνευσης μεταβολών, οι οποίοι έχουν ως στόχο τον εντοπισμό μη φυσιολογικών αποκλίσεων στις τρέχουσες τιμές σε σχέση με τις τιμές που προέκυψαν σε προηγούμενα βήματα. Οι αλγόριθμοι ανίχνευσης μεταβολών μπορούν να ταξινομηθούν σε δύο κατηγορίες, η ανίχνευση μεταβολών μίας μεταβλητής και η ανίχνευση μεταβολών πολλαπλών μεταβλητών.

Οι αλγόριθμοι οι οποίοι ανήκουν στην κατηγορία ανίχνευσης μεταβολών μίας μεταβλητής λαμβάνουν υπόψη κάθε ροή αισθητήρων ξεχωριστά και κάνουν ανίχνευση πιθανών ανωμαλιών μέσα από μία ακολουθιακή ανάλυση χρονοσειρών. Οι αλγόριθμοι οι οποίοι ανήκουν στην κατηγορία ανίχνευσης μεταβολών πολλαπλών μεταβλητών εκμεταλλεύονται αυτοπαλίνδρομα μοντέλα πολλαπλών μεταβλητών για την αναπαράσταση κάθε διανύσματος πλαισίου ως ένα γραμμικό άθροισμα της προηγούμενης συμπεριφοράς. Στη συνέχεια, ο στόχος απόκτησης μίας δυαδικής τιμής η οποία υποδεικνύει την μεταβολή ή μη μεταβολή για κάποια συγκεκριμένη μεταβλητή, δηλαδή μία ροή αισθητήρων, ανάγεται σε μία λειτουργία ελέγχου κατωφλίου μεταξύ του μελλοντικού εκτιμώμενου διανύσματος και του πραγματικού διανύσματος.

Οι μέθοδοι ανίχνευσης μεταβολών λαμβάνουν υπόψη τη χρονική σειρά των τιμών των μετρήσεων και κάνουν αναζήτηση για χρονικά σημεία στα οποία οι στατιστικές ιδιότητες των μετρήσεων αλλάζουν απότομα. Σύμφωνα με το [2] η λέξη απότομα συγκεκριμενοποιείται ως «σε άμεσο χρόνο ή τουλάχιστον πολύ γρήγορα εάν λάβουμε υπόψη την περίοδο δειγματοληψίας των μετρήσεων». Οι παρακολουθούμενες στατιστικές ιδιότητες θεωρείται ότι παρουσιάζουν καμία ή πολύ μικρή απόκλιση στις χρονικές στιγμές στις οποίες δεν παρατηρείται κάποια μεταβολή. Λαμβάνοντας υπόψη τις προαναφερθείσες συνθήκες, μπορεί να γίνει ανίχνευση ακόμα και μικρών μεταβολών με αρκετά μεγάλη πιθανότητα. Η πιθανότητα ανίχνευσης μπορεί να είναι ακόμα πιο μεγάλη εάν αυτές οι μεταβολές είναι επίμονες για κάποιο μεγάλο χρονικό διάστημα.

Οι μέθοδοι ανίχνευσης μεταβολών στις πλείστες των περιπτώσεων λειτουργούν χωρίς κάποια υπόθεση ότι οι παρακολουθούμενες μεταβλητές περιγράφονται από κάποια συγκεκριμένη κατανομή. Με άλλα λόγια, οι μέθοδοι ανίχνευσης μεταβολών συνήθως είναι μη παραμετρικές. Ακόμα ένα χαρακτηριστικό των μεθόδων ανίχνευσης

μεταβολών αποτελεί η ανίχνευση των μεταβολών σε πολύ σύντομο χρονικό διάστημα ή ακόμα και άμεσα. Επίσης, η πληροφορία μεγέθους κάποιας μεταβολής στις περισσότερες περιπτώσεις δεν είναι κάτι μετρήσιμο ή απαραίτητο.

Ο σχεδιασμός διαδικασιών ανίχνευσης απότομων μεταβολών αποτελείται από δύο μεγάλες υποδιαδικασίες. Η πρώτη υποδιαδικασία είναι προαιρετική και περιλαμβάνει μία επεξεργασία των αρχικών δεδομένων έτσι ώστε οι τελικές τιμές του συνόλου δειγμάτων να μην αποκλίνουν κατά πολύ, από μία αρχική τιμή, από μετρικές όπως ο μέσος όρος, η απόκλιση και άλλα, όταν δεν παρατηρείται μεταβολή. Η αρχική τιμή μπορεί να είναι μηδενική ή κάποια άλλη κατάλληλη τιμή. Σε αυτή την υποδιαδικασία οι τελικές τιμές του συνόλου δειγμάτων αποκλίνουν σε σημαντικό βαθμό από την προαναφερθείσα τιμή αναφοράς όταν παρατηρείται κάποια μεταβολή. Η δεύτερη διαδικασία περιλαμβάνει την ανάπτυξη αλγορίθμων που ανήκουν στην κατηγορία των στατιστικών μεθόδων. Οι αλγόριθμοι αυτοί πρέπει να είναι ικανοί για ανίχνευση των απότομων μεταβολών στο σύνολο δειγμάτων και τις ακριβείς χρονικές στιγμές κατά τις οποίες εμφανίστηκαν.

Στις επόμενες παραγράφους γίνεται μία περιγραφή των αλγορίθμων ανίχνευσης μεταβολών που εξετάζονται σε αυτή την εργασία για τον εντοπισμό πραγματικού χρόνου των συμβάντων σε ροές αισθητήρων.

2.1 Μετρικές απόδοσης αλγορίθμων ανίχνευσης συμβάντων

Ο στόχος ενός αλγορίθμου ανίχνευσης απότομων μεταβολών είναι η ανίχνευση μίας απότομης μεταβολής στην κατανομή πιθανοτήτων μίας ακολουθίας από τυχαίες παρατηρήσεις. Ο ιδανικός αλγόριθμος ανίχνευσης απότομων μεταβολών δεν ανιχνεύει μία απότομη μεταβολή, μέχρι να συμβεί μία απότομη μεταβολή, και όταν η απότομη μεταβολή συμβεί, η ανίχνευση γίνεται σε άμεσο χρόνο. Ωστόσο, λόγω της στοχαστικής διακύμανσης, η ιδανική αυτή περίπτωση δεν είναι εφικτό να συμβεί.

Στην πράξη, υπάρχουν χρονικές στιγμές στις οποίες ο αλγόριθμος κάνει ανίχνευση απότομης μεταβολής, όταν αυτή δεν έχει συμβεί. Η περίπτωση αυτή ονομάζεται ψευδής συναγερμός (false alarm). Επίσης, στις περιπτώσεις απότομων μεταβολών, υπάρχει κάποια καθυστέρηση στην ανίχνευση της αλλαγής. Συνοπτικά, υπάρχουν δύο τυπικές μετρικές απόδοσης, $MΔE_0$ και $MΔE_1$, όπου $MΔE$ ορίζεται ως η μέση διάρκεια εκτέλεσης [32]. Η μετρική $MΔE_0$ ορίζεται ως ο μέσος χρόνος μεταξύ ψευδών συναγερμών, ενώ η μετρική $MΔE_1$ ορίζεται ως ο χρόνος μέσης καθυστέρησης μεταξύ χρόνου που συμβαίνει απότομη μεταβολή και του χρόνου που γίνεται η ανίχνευση από τον αλγόριθμο. Σε ιδανική

περίπτωση, ένας αλγόριθμος απότομων μεταβολών έχει υψηλή τιμή $M\Delta E_0$ και χαμηλή τιμή $M\Delta E_1$. Ωστόσο, συντονίζοντας τις παραμέτρους ενός αλγορίθμου για επίτευξη επιθυμητής τιμής ενός από τα μεγέθη $M\Delta E_0$ και $M\Delta E_1$ έχει αρνητική επίπτωση στο άλλο μέγεθος.

2.2 Γενικός αλγόριθμος διαγραμμάτων ελέγχου

Μία συνήθης μέθοδος στατιστικής ανίχνευσης μεταβολών σε άμεσο χρόνο είναι η μέθοδος διαγραμμάτων ελέγχου [33]. Σε ένα διάγραμμα ελέγχου, η μέση τιμή και η μεταβλητότητα μίας παρακολουθούμενης μεταβλητής περιγράφονται με τα μεγέθη κεντρικού άξονα, ανώτατο όριο ελέγχου και κατώτατο όριο ελέγχου. Γίνεται ανίχνευση μίας μεταβολής εάν η τιμή της μεταβλητής σε κάποια χρονική στιγμή υπερβαίνει κάποιο από τα όρια ελέγχου, το ανώτατο όριο ελέγχου ή το κατώτατο όριο ελέγχου. Η ανίχνευση μεταβολής μπορεί να εκφραστεί με έναν έλεγχο υποθέσεων με μηδενική υπόθεση Y_0 να περιγράφει καμία σημαντική μεταβολή και η συμπληρωματική υπόθεση Y_1 να περιγράφει κάποια σημαντική μεταβολή.

Η ιδέα διαγραμμάτων ελέγχου περιγράφηκε για πρώτη φορά με αρχικό κίνητρο την ανίχνευση αλλαγής στις διαδικασίες παραγωγής, για σκοπούς ελέγχου ποιότητας. Ένα διάγραμμα ελέγχου αποτελείται από σημεία $y_1, y_2, \dots$ τα οποία αντιπροσωπεύουν στατιστικά όρια και όρια ελέγχου α και δ, όπου ισχύει $\alpha < \delta$. Όταν ισχύει $y_\kappa \in (\alpha, \delta)$, η διαδικασία βρίσκεται σε κατάσταση ελέγχου, ενώ όταν ισχύει $y_\kappa \notin (\alpha, \delta)$, η διαδικασία βρίσκεται σε κατάσταση εκτός ελέγχου. Μία ακολουθία από στατιστικές συμβολίζεται με y_κ, συμβολισμός διαφορετικός από το συμβολισμό x_κ, ο οποίος αντιπροσωπεύει μία ακολουθία από παρατηρηθείσες τιμές. Το α ονομάζεται ανώτατο όριο ελέγχου και το δ ονομάζεται κατώτατο όριο ελέγχου. Το $\hat{\tau}$ ονομάζεται το σημείο αλλαγής της ροής δεδομένων, εάν ισχύει $y_{\hat{\tau}} \notin (\alpha, \delta)$, αλλά $y_t \in (\alpha, \delta)$ για όλα τα $t < \hat{\tau}$. Στον παραπάνω συμβολισμό το τ συμβολίζει ένα πραγματικό σημείο αλλαγής.

Ο αλγόριθμος διαγραμμάτων ελέγχου διατηρεί έναν κεντρικό άξονα, ο οποίος αντιπροσωπεύει τη μέση τιμή της παρακολουθούμενης μεταβλητής κάτω από κανονικές συνθήκες. Επίσης διατηρεί τιμές ανώτατου ορίου ελέγχου και κατώτατου ορίου ελέγχου, τα οποία όρια έχουν τιμές πάνω και κάτω από την τιμή κεντρικού άξονα αντίστοιχα, και καθορίζουν το εύρος της κατάστασης που χαρακτηρίζεται από κανονική μεταβλητότητα ή μεταβλητότητα εντός ελέγχου. Η συνάρτηση ελέγχου μεταβλητότητας κάνει μετάβαση

στην κατάσταση μη κανονικής μεταβλητότητας ή μεταβλητότητας εκτός ελέγχου εάν η μετρηθείσα τιμή της μεταβλητής είναι εκτός των προκαθορισμένων ορίων.

2.3 Διαγράμματα ελέγχου αλγορίθμου συσσωρευτικού αθροίσματος

Η μέθοδος διαγραμμάτων αλγορίθμου συσσωρευτικού αθροίσματος είναι ένα από τα πρώτα εργαλεία που χρησιμοποιήθηκαν στην ανίχνευση απότομων μεταβολών. Προτάθηκε για πρώτη φορά το 1954 [11] και έχει μελετηθεί ευρέως στη βιβλιογραφία. Η μέθοδος είναι εύκολα διαχειρίσιμη και χρήσιμη για ανίχνευση των θέσεων των σημείων μεταβολής. Συγκεκριμένα, έχει χρησιμοποιηθεί για έλεγχο μεταβολής στις συναρτήσεις μέσης τιμής, απόκλισης και κατανομής.

Ένα πλεονέκτημα της μεθόδου έγκειται στο γεγονός ότι οι συναρτήσεις μέσης τιμής, απόκλισης και κατανομής εκφράζονται ως το άθροισμα ανεξάρτητων και πανομοιότυπα κατανεμημένων τυχαίων μεταβλητών. Έχει επίσης το πλεονέκτημα να λαμβάνει υπόψη το ιστορικό της υπό διερεύνηση σειράς και είναι ικανό να ανιχνεύσει αποτυχία μοντέλου πιο γρήγορα όταν το σφάλμα πρόβλεψης είναι σχετικά μικρό. Το διάγραμμα αλγορίθμου συσσωρευτικού αθροίσματος ενσωματώνει άμεσα όλη την πληροφορία στην ακολουθία του δείγματος με τη γραφική παράσταση των συσσωρευμένων αθροισμάτων των αποκλίσεων των τιμών του συνόλου δειγμάτων από την τιμή του κεντρικού άξονα. Το διάγραμμα μπορεί να κατασκευαστεί τόσο για μεμονωμένες παρατηρήσεις, όσο και για τους μέσους όρους των λογικών υποομάδων του συνόλου δειγμάτων.

2.3.1 Ο αλγόριθμος συσσωρευτικού αθροίσματος και θεωρία ελέγχου αποφάσεων

Ο αλγόριθμος συσσωρευτικού αθροίσματος [11] [13] βασίζεται στο γεγονός ότι το μέγεθος $\sigma_t = \sigma(y_1, \dots, y_t)$ έχει αρνητική τάση τιμών σε κανονικές συνθήκες και θετική τάση τιμών μετά από μία μεταβολή. Η συνάρτηση απόφασης α_t συγκρίνει την αύξηση του σ_t από την ελάχιστη τιμή του με ένα κατώφλι κ:

$$\alpha_t = \sigma_t - min_{1 \le i \le t} s_i = \max(0, \sigma(y_t) + a_{t-1}) = [a_{t-1} + \sigma(y_t)]^+ \ge \kappa \qquad \text{όπου } \alpha_0 = 0$$

Γίνεται ανίχνευση ενός συμβάντος που περιγράφει μεταβολή εάν η συνάρτηση α_t ξεπερνά την τιμή κατωφλίου κ. Στην περίπτωση αυτή, εάν ο αλγόριθμος συνεχίζει και σε επόμενες χρονικές στιγμές, ο αλγόριθμος επανεκκινεί με τιμή μηδέν στη συνάρτηση α_t.

Ο αλγόριθμος συσσωρευτικού αθροίσματος μπορεί να περιγραφεί και με τη θεωρία ελέγχου υποθέσεων. Σε μια τέτοια περιγραφή, το διάγραμμα ελέγχου αλγορίθμου συσσωρευτικού αθροίσματος εκτελεί με επαναληπτική συμπεριφορά έναν Ακολουθιακό Έλεγχο Λόγου Πιθανοφανειών, στον οποίο κάθε απόφαση λαμβάνει υπόψη όσες διαδοχικές παρελθοντικές παρατηρήσεις είναι απαραίτητο για να γίνει η αποδοχή της κατάστασης Υ_0 ή Υ_1. Ο αλγόριθμος επανεκκινά τη διαδικασία απόφασης με τη μέθοδο Ακολουθιακού Ελέγχου Λόγου Πιθανοφανειών εάν έχει γίνει αποδεκτή η κατάσταση Υ_0. Σε αντίθετη περίπτωση, εάν γίνει αποδεκτή η κατάσταση Υ_1, γίνεται σηματοδότηση ανίχνευσης μεταβολής και ο αλγόριθμος σταματά. Η τιμή κατωφλίου κ προσφέρει μία εξισορρόπηση μεταξύ του μέσου χρόνου καθυστέρησης ανίχνευσης και του μέσου χρόνου μεταξύ λανθασμένων ανιχνεύσεων.

Μία τυπική στατιστική για ανίχνευση θετικών αποκλίσεων από την τιμή κεντρικού άξονα είναι $u^+(y) = y - (\mu_0 + K)$, όπου το K ονομάζεται τιμή αναφοράς. Για ανίχνευση και αρνητικών αποκλίσων, μία δεύτερη στατιστική είναι απαραίτητη $u^-(y) = (\mu_0 - K) - y$. Οι συναρτήσεις ανίχνευσης μεταβολών ορίζονται:

$$\alpha_t^+ = [a_{t-1}^+ + y_t - (\mu_0 + K)]^+ \geq \kappa$$

$$\alpha_t^- = [a_{t-1}^- + (\mu_0 - K) - y_t]^+ \geq \kappa$$

Τυπικές τιμές των K και κ είναι $K = \frac{\sigma}{2}$ και $\kappa = 4\sigma$ ή $\kappa = 5\sigma$ όπου σ είναι η τυπική απόκλιση της Y_t.

2.3.2 Ο αλγόριθμος συσσωρευτικού αθροίσματος CUSUM

Ο αλγόριθμος συσσωρευτικού αθροίσματος [11] έχει ως στόχο την ανίχνευση, σε πραγματικό χρόνο, μίας μεταβολής στην κατανομή μίας χρονοσειράς σε σχέση με μία τιμή - στόχο. Συγκεκριμένα, αρχικά γίνεται η υπόθεση μίας μονοδιάστατης χρονοσειράς x_t η οποία αποτελείται από τιμές δεδομένων, οι οποίες συλλέχθηκαν με την πάροδο του χρόνου, και μίας τιμής - στόχος μ για αυτή τη ροή δεδομένων. Ο αλγόριθμος συσσωρευτικού αθροίσματος περιλαμβάνει τον υπολογισμό θετικών μεταβολών Θ και αρνητικών μεταβολών A στη χρονοσειρά x_t συσσωρευτικά με την πάροδο του χρόνου

και συγκρίνει αυτές τις μεταβολές με ένα θετικό κατώφλι $κατώφλι^+$ και ένα αρνητικό κατώφλι $κατώφλι^-$. Κάθε φορά που οι τιμές κατωφλίων ξεπερνούνται, γίνεται μία αναφορά μεταβολής μέσω του σήματος άνω ανίχνευσης σ^+ και του σήματος κάτω ανίχνευσης σ^-, ενώ τα συσσωρευτικά αθροίσματα παίρνουν μηδενική τιμή. Προκειμένου να αποφευχθεί η ανίχνευση μη απότομων μεταβολών ή αργών μετατοπίσεων, ο αλγόριθμος λαμβάνει υπόψη παραμέτρους ανεκτικότητας για θετικές μεταβολές α^+ και για αρνητικές μεταβολές α^-.

Οι παράμετροι εισόδου για τον αλγόριθμο συσσωρευτικού αθροίσματος είναι η τιμή στόχος μ, η τιμή άνω ανεκτικότητας α^+, η τιμή κάτω ανεκτικότητας α^-, η τιμή άνω κατωφλίου $κατώφλι^+$ και η τιμή κάτω κατωφλίου $κατώφλι^-$. Οι παράμετροι εξόδου είναι το σήμα άνω ανίχνευσης σ^+ και το σήμα κάτω ανίχνευσης σ^-. Στη συνέχεια γίνεται παρουσίαση του αλγορίθμου συσσωρευτικού αθροίσματος.

Είσοδος: τιμή στόχος μ, τιμή άνω ανεκτικότητας α^+, τιμή κάτω ανεκτικότητας α^-, τιμή άνω κατωφλίου $κατώφλι^+$, τιμή κάτω κατωφλίου $κατώφλι^-$

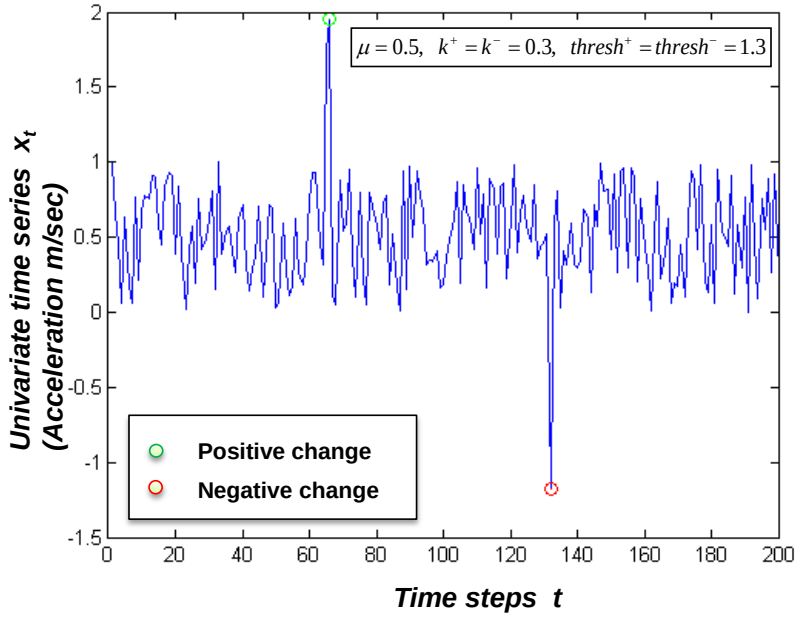
Έξοδος: σήμα άνω ανίχνευσης σ^+, σήμα κάτω ανίχνευσης σ^-

```
1:  Θ ← 0;
2:  A ← 0;
3:  t ←1;
4:  while ( true )
5:      σ+ ← 0;
6:      σ− ← 0;
7:      Θ ← max (0, x_t − (μ + α+) + Θ);
8:      A ← min (0, x_t − (μ − α−) + A);
9:      if (Θ > κατώφλι+) then
10:         σ+ ← 1;
11:         Θ ← 0;
12:         A ← 0;
13:     end
14:     if (A < κατώφλι−) then
15:         σ+ ← 1;
16:         Θ ← 0;
17:         A ← 0;
18:     end
19:     t ←t +1;
end
```

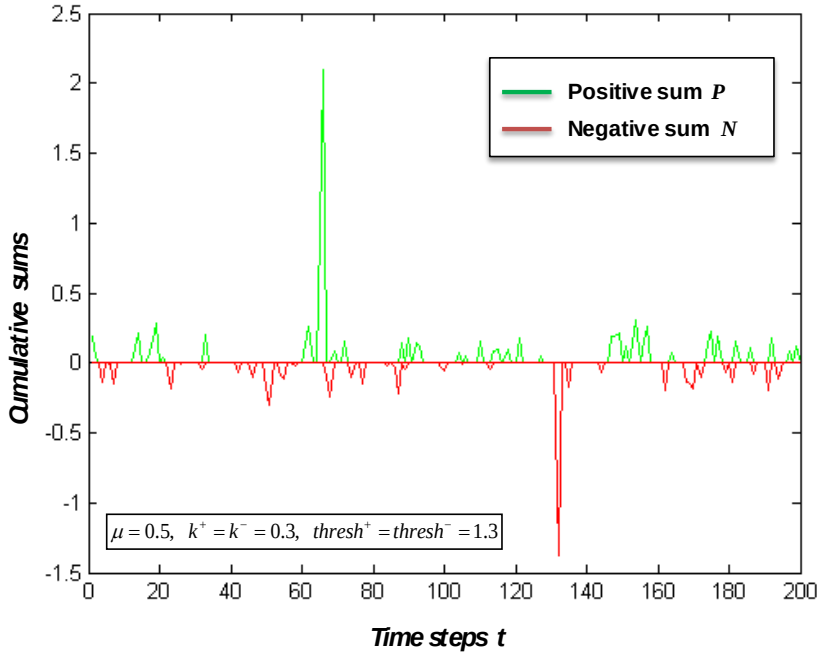
Αλγόριθμος 1: Αλγόριθμος Συσσωρευτικού Αθροίσματος

Ο αλγόριθμος υποθέτει ότι η χρονοσειρά που καταφθάνει ακολουθεί κανονική κατανομή. Προκειμένου ο αλγόριθμος να λειτουργήσει σωστά, πρέπει να γίνει συντονισμός των παραμέτρων ανεκτικότητας και κατωφλίου με τρόπο που να καθορίζει τι είναι μία πραγματική μεταβολή για κάποια συγκεκριμένη χρονοσειρά. Ο συντονισμός αυτός μπορεί να εκτελεστεί ακολουθώντας κάποια βήματα [34]. Αρχικά, γίνεται έναρξη της διαδικασίας με μεγάλες τιμές των $κατώφλι^{+}$ και $κατώφλι^{-}$. Στη συνέχεια, επιλέγονται οι παράμετροι άνω ανεκτικότητας α^{+} και κάτω ανεκτικότητας α^{-} στο μέσο της αναμενόμενης μεταβολής ή γίνεται προσαρμογή τους έτσι ώστε οι μετρικές Θ και A να έχουν μηδενική τιμή περισσότερες από το ένα δεύτερο φορές. Οι τιμές των $κατώφλι^{+}$ και $κατώφλι^{-}$ ρυθμίζονται έτσι ώστε να επιτυγχάνεται ο απαιτούμενος αριθμός λανθασμένων συναγερμών ή ο απαιτούμενος χρόνος καθυστέρησης. Εάν το απαιτούμενο είναι ταχύτερη ανίχνευση, γίνεται μείωση των τιμών άνω ανεκτικότητας α^{+} και κάτω ανεκτικότητας α^{-}. Εάν είναι επιθυμητός μικρότερος αριθμός από λανθασμένους συναγερμούς ή γίνεται ανίχνευση μεταβολών οι οποίες δεν βγάζουν νόημα, γίνεται αύξηση των τιμών άνω ανεκτικότητας α^{+} και κάτω ανεκτικότητας α^{-}.

Στις Εικόνες 3 και 4 απεικονίζεται ένα παράδειγμα του αλγορίθμου συσσωρευτικού αθροίσματος πάνω σε μία ροή αισθητήρων, όπου γίνεται ανίχνευση δύο μεταβολών, θετικής και αρνητικής. Η τιμή στόχος για x_i ορίζεται σε $\mu = 0{,}5$, οι τιμές ανεκτικότητας ορίζονται σε $\alpha^{+} = \alpha^{-} = 0{,}3$ και οι τιμές κατωφλίου προσδιορίζονται σε $κατώφλι^{+} = κατώφλι^{-} = 1{,}3$. Συγκεκριμένα, η Εικόνα 3 παρουσιάζει τα αρχικά δεδομένα αισθητήρων και τα χρονικά βήματα στα οποία ανιχνεύεται μεταβολή από τον αλγόριθμο. Η Εικόνα 4 απεικονίζει τα συσσωρευτικά αθροίσματα των θετικών και αρνητικών μεταβολών σε συνάρτηση με το χρόνο για x_t. Στην περίπτωση δεδομένων αισθητήρων πολλαπλών μεταβλητών, ο αλγόριθμος συσσωρευτικού αθροίσματος πρέπει να εφαρμόζεται σε κάθε μεταβλητή ξεχωριστά.



Εικόνα 3: Αυθεντική ροή αισθητήρων καταμέτρησης επιτάχυνσης μέσω MPU και ανίχνευση μεταβολών με το αλγόριθμο συσσωρευτικού αθροίσματος



Εικόνα 4: Συσσωρευτικά αθροίσματα θετικών και αρνητικών μεταβολών

2.3.3 Μονόπλευρος και αμφίπλευρος αλγόριθμος συσσωρευτικού αθροίσματος

Οι αλγόριθμος συσσωρευτικού αθροίσματος μπορεί να διαχωριστεί σε δύο τύπους ή παραλλαγές [13] [35]. Ο πρώτος τύπος είναι ο μονόπλευρος αλγόριθμος συσσωρευτικού αθροίσματος, ο οποίος μπορεί να χρησιμοποιηθεί όταν οι τιμές του μέσου όρου πριν και

μετά από απότομη μεταβολή είναι γνωστές εκ των προτέρων. Ο δεύτερος τύπος είναι ο αμφίπλευρος CUSUM, ο οποίος χρησιμοποιείται όταν το μέγεθος της μεταβολής δεν είναι γνωστό εκ των προτέρων.

2.3.3.1 Μονόπλευρος αλγόριθμος συσσωρευτικού αθροίσματος

Η συνάρτηση $\lambda_n = \sum_{i=1}^{n} z_i$ όπου

$$z_i = \ln \frac{P_{\theta_1}(y_i)}{P_{\theta_0}(y_i)} = \frac{\mu_1 - \mu_o}{\sigma^2}\left(y_i - \frac{\mu_0 + \mu_1}{2}\right) = \frac{\beta}{\sigma}\left(y_i - \frac{\mu_0 + \mu_1}{2}\right) = \frac{\beta}{\sigma}(y_i - \mu_0 - \frac{v}{2})$$

είναι ο λογαριθμικός λόγος πιθανοφανειών για τις παρατηρηθείσες τιμές από y_1 έως y_n, και το μέγεθος μεταβολής ορίζεται από $v = \mu_1 - \mu_0$ και ισχύει $\beta = \frac{\mu_1 - \mu_0}{\sigma}$. Στο n σημείο δειγματοληψίας, η συνάρτηση απόφασης έχει συσσωρευτικό χαρακτήρα και εκφράζεται ως:

$$\alpha_n = \frac{\beta}{\sigma}\sum_{i=1}^{n}(y_i - \mu_0 - \frac{v}{2}).$$

Στην περίπτωση θετικής μεταβολής στο μέσο όρο, η τυπική συμπεριφορά της συνάρτησης απόφασης παρουσιάζει μία αρνητική μετατόπιση πριν την μεταβολή και μία θετική μετατόπιση μετά την μεταβολή. Στην περίπτωση αρνητικής μεταβολής στο μέσο όρο, η τυπική συμπεριφορά της συνάρτησης απόφασης παρουσιάζει μία θετική μετατόπιση πριν τη μεταβολή και μία αρνητική μετατόπιση μετά τη μεταβολή.

Στην περίπτωση του μονόπλευρου αλγορίθμου συσσωρευτικού αθροίσματος, η βασική υπόθεση είναι ότι έχει παρουσιαστεί μία μεταβολή γνωστού μεγέθους στο μέσο όρο, μεταβολή που μπορεί να είναι θετική ή αρνητική. Στην περίπτωση θετικής μεταβολής στο μέσο όρο, η σχετική πληροφορία μεταβολής προκύπτει από τη διαφορά μεταξύ του λογαριθμικού λόγου πιθανοφανειών και του μέχρι τώρα ελαχίστου των τιμών δειγμάτων του συνόλου. Ο κανόνας απόφασης στην περίπτωση αυτή είναι η σύγκριση, σε κάθε χρονική στιγμή, της διαφοράς που προαναφέρθηκε με ένα προκαθορισμένο κατώφλι. Συνεπώς, εμφανίζεται συναγερμός απότομης μεταβολής εάν ισχύει

$$\sigma_n = \alpha_n - min_{0\le t\le n} a_t \ge \lambda \text{ όπου } \alpha_t = (\alpha_{t-1} + y_t - \mu_0 - \frac{v}{2})^+, \alpha_0 = 0$$

όπου σ_n είναι η συνάρτηση απόφασης αλγορίθμου συσσωρευτικού αθροίσματος.

2.3.3.2 Αμφίπλευρος αλγόριθμος συσσωρευτικού αθροίσματος

Μία άλλη περίπτωση ανίχνευσης μεταβολής χρησιμοποιώντας τον αλγόριθμο συσσωρευτικού αθροίσματος είναι η περίπτωση στην οποία το μέγεθος της μεταβολής δεν είναι γνωστό. Στην περίπτωση αυτή, το μέγεθος της μεταβολής εκφράζεται ως ένα προκαθορισμένο ελάχιστο μέγεθος μεταβολής, το οποίο μπορεί να είναι $\mu_1^+ = \mu_0 + v$ στην περίπτωση της θετικής μετατόπισης στο μέσο όρο και $\mu_1^- = \mu_0 - v$ στην περίπτωση της αρνητικής μετατόπισης στο μέσο όρο.

Στον αμφίπλευρο έλεγχο αλγορίθμου συσσωρευτικού αθροίσματος, είναι χρήσιμο να γίνεται εκτέλεση δύο παράλληλων ελέγχων, όπου ο πρώτος εμφανίζει συναγερμό απότομης μεταβολής στην περίπτωση αύξησης στο μέσο όρο και ο δεύτερος στην περίπτωση μείωσης στο μέσο όρο. Συνεπώς, εμφανίζεται συναγερμός απότομης μεταβολής εάν ισχύει

$$\sigma_n^+ = a_n^+ - min_{0\le t\le n} a_t^+ \ge \lambda$$

$$\sigma_n^- = max_{0\le t\le n} a_t^- - a_n^- \ge \lambda$$

όπου $a_t^- = (a_{t-1}^- - y_t + \mu_0 - \frac{v}{2})^+, a_0^+ = a_0^- = 0$

Η προκύπτουσα χρονική στιγμή εμφάνισης συναγερμού απότομης μεταβολής δίνεται από:

$$t_{συναγερμός} = \min\{t \ge 1 : (a_t^+ \ge \lambda) \cup (a_t^- \ge \lambda)\}$$

Η παράμετρος μεγέθους μεταβολής, εφόσον δεν είναι εκ των προτέρων γνωστή, μπορεί να έχει διάφορες μορφές, ανάλογα με τα χαρακτηριστικά της εφαρμογής. Στις περισσότερες εφαρμογές η διαθέσιμη πληροφορία όσον αφορά το μέγεθος αυτό είναι πολύ μικρή. Το μέγεθος μεταβολής μπορεί να καθοριστεί ως το μικρότερο δυνατό μέγεθος μεταβολής, το πιο πιθανό μέγεθος μεταβολής ή το μέγεθος για το οποίο δεν εμφανίζεται συναγερμός για ένα ανεκτό αριθμό περιπτώσεων, ενώ υπήρχε απότομη μεταβολή.

2.3.4 Προσέγγιση τιμών μέσου όρου και τυπικής απόκλισης

Ο λανθασμένος καθορισμός τιμών μέσου όρου και τυπικής απόκλισης των προς διερεύνηση δεδομένων είναι πιθανό να οδηγήσει σε ένα μη επιθυμητό αριθμό από

λανθασμένους συναγερμούς. Σε κάποιες εφαρμογές, τα δεδομένα ελέγχου μπορούν να χρησιμοποιηθούν για τον υπολογισμό της μέσης τιμής και της τυπικής απόκλισης. Στη συνέχεια, οι αρχικές προσεγγίσεις των παραμέτρων μπορούν να χρησιμοποιηθούν για τη βαθμονόμηση του αλγορίθμου συσσωρευτικού αθροίσματος.

Στην περίπτωση που δεν υπάρχουν διαθέσιμα δεδομένα ελέγχου στο σύνολο δεδομένων, χρησιμοποιείται ένας αλγόριθμος συσσωρευτικού αθροίσματος ο οποίος ξεκινάει χωρίς κάποια εκ των προτέρων πληροφορία, όσον αφορά τις τιμές μέσου όρου και τυπικής απόκλισης. Ο αλγόριθμος αυτός υπολογίζει τις τιμές μέσου όρου και τυπικής απόκλισης σε πραγματικό χρόνο, χρησιμοποιώντας τα δεδομένα που έχουν συγκεντρωθεί μέχρι την τρέχουσα χρονική στιγμή. Σε ένα τέτοιο σύστημα, οι τιμές μέσου όρου και τυπικής απόκλισης ενημερώνονται ακολουθιακά και σε πραγματικό χρόνο.

2.3.5 Προσέγγιση τιμών αναμενόμενου μεγέθους μετατόπισης και κατωφλίου

Για την επίτευξη του βέλτιστου αλγορίθμου συσσωρευτικού αθροίσματος, πρέπει να γίνει προσέγγιση της αναμενόμενης τιμής μετατόπισης με τέτοιο τρόπο ώστε να ελαχιστοποιούνται οι αριθμοί χαμένων απότομων μεταβολών και λανθασμένων συναγερμών. Το κριτήριο βελτιστοποίησης είναι η επιλογή τιμής μετατόπισης, η οποία προσφέρει ανίχνευση απότομης μεταβολής με μικρό χρόνο καθυστέρησης και μικρό αριθμό λανθασμένων συναγερμών.

Τυπικά, το μέγεθος αναμενόμενης μετατόπισης μπορεί να οριστεί της τάξης ενός ή δύο τυπικών αποκλίσεων των δεδομένων ελέγχου. Εκτός από το μέγεθος αναμενόμενης μετατόπισης, η επιλογή τιμής κατωφλίου μπορεί επίσης να επηρεάσει το χρόνο καθυστέρησης ανίχνευσης απότομης μεταβολής. Τυπικά, αρχικά γίνεται μία προσέγγιση της τιμής μεγέθους αναμενόμενης μεταβολής και στη συνέχεια καθορισμός της τιμής κατωφλίου.

2.3.6 Επιλογή κριτηρίων

Το τυπικό μέγεθος απόδοσης για ανίχνευση απότομης μεταβολής σε πραγματικό χρόνο είναι ο χρόνος καθυστέρησης ανίχνευσης, ο οποίος πρέπει να ελαχιστοποιείται για ένα σταθερό ποσοστό λανθασμένων συναγερμών. Μπορεί επίσης να γίνει ορισμός των

κριτηρίων, τα οποία μπορούν να χρησιμοποιηθούν για προσέγγιση βέλτιστων χρόνων διακοπής στην ανίχνευση απότομων μεταβολών. Τα κριτήρια πρέπει να ευνοούν ανίχνευση απότομων μεταβολών με ελάχιστο χρόνο καθυστέρησης και μικρό αριθμό λανθασμένων συναγερμών. Ο χρόνος καθυστέρησης ανίχνευσης απότομων μεταβολών είναι επιθυμητός για το λόγο ότι, η διαφορά μεταξύ των χρόνων πραγματικής απότομης μεταβολής και εμφάνισης συναγερμού αυτής μπορεί να οδηγήσει σε εσφαλμένη ερμηνεία των χαρακτηριστικών του προς διερεύνηση μεγέθους.

Από τα προαναφερθέντα εμφανίζεται η ανάγκη για εξισορρόπηση μεταξύ των μεγεθών μέσου χρόνου μεταξύ λανθασμένων συναγερμών και του χρόνου καθυστέρησης ανίχνευσης απότομης μεταβολής. Τα δύο προαναφερθέντα μεγέθη αυξάνονται, όσο ο αλγόριθμος είναι λιγότερο ευαίσθητος σε υψηλές συχνότητες. Μία σχετική εξισορρόπηση μεγεθών, η οποία είναι επίσης επιθυμητή, είναι η εξισορρόπηση μεταξύ των μεγεθών αποδοτικότητας και πολυπλοκότητας. Όταν το σχεδιασμένο σύστημα ανίχνευσης περιλαμβάνει σε κάθε συναγερμό απότομης μεταβολής μία επεξεργασία, η οποία είναι χρονοβόρα ή και ακριβή όσον αφορά τις πράξεις υπολογισμού, ή και αναδιαμόρφωση ελέγχου, ο αριθμός λανθασμένων συναγερμών είναι πιο μεγάλος. Σε κάποια συστήματα, είναι χρήσιμο να υπάρχει μείωση της πολυπλοκότητας υπολογισμού, χωρίς επηρεασμό της αποδοτικότητας του αλγορίθμου. Ένα παράδειγμα για την επίτευξη του στόχου αυτού είναι η χρήση πιθανού πλεονασμού δεδομένων.

2.4 Διαγράμματα ελέγχου αλγορίθμου Shewhart

Τα διαγράμματα ελέγχου αλγορίθμου Shewhart [25] καθορίζουν ένα μέγεθος κεντρικού άξονα, όπως επίσης και τα μεγέθη ανώτατου ορίου ελέγχου AOE και κατώτατου ορίου ελέγχου KOE. Τα μεγέθη αυτά καθορίζονται με βάση κάποια μετρική, η οποία προσδιορίζεται από N παρατηρηθείσες τιμές της μεταβλητής $y_{(j-1)N+1}, \dots, y_{jN}$. Ένα παράδειγμα τέτοιας μετρικής είναι η μέση τιμή $\bar{y}_j$, η οποία είναι κατάλληλη για ανίχνευση μεταβολών στο μέσο όρο:

$$\bar{y}_j = \frac{1}{N}\sum_{t=(j-1)N+1}^{jN} y_t, \text{ όπου } j = 1,2, \dots$$

Εάν οι παρατηρηθείσες τιμές είναι ανεξάρτητες και ομοιόμορφα κατανεμημένες με μέσο μ_0 και τυπική απόκλιση σ^2, η μετρική $\bar{y}_j$ είναι μία μετρική εκτίμησης του μ_0 με απόκλιση $\frac{\sigma^2}{N}$. Ως εκ τούτου, το ανώτατο όριο ελέγχου και το κατώτατο όριο ελέγχου ορίζονται ως εξής:

$$AOE = \mu_0 + \frac{L\sigma}{\sqrt{N}}$$

$$KOE = \mu_0 - \frac{L\sigma}{\sqrt{N}}$$

όπου L είναι η παράμετρος συντονισμού. Ενεργοποιείται ένας συναγερμός ανίχνευσης μεταβολής, εάν η τιμή $\overline{y_j}$ ξεπερνάει την τιμή ανώτατου όριου ελέγχου ή την τιμή κατώτατου όριου ελέγχου.

2.4.1 Ο αλγόριθμος διαγραμμάτων ελέγχου Shewhart και θεωρία ελέγχου υποθέσεων

Οι μέθοδοι ανίχνευσης οριακών τιμών αρχικά παρουσιάστηκαν στο πεδίο του ποιοτικού ελέγχου, και συγκεκριμένα στα συστήματα συνεχούς ελέγχου. Στη συνέχεια περιγράφονται οι μαθηματικές έννοιες, οι οποίες χαρακτηρίζουν τα συστήματα αυτά [13] [36]. Γίνεται υπόθεση K τιμών δειγμάτων ενός συνόλου δεδομένων με σταθερό μέγεθος N. Στο τέλος του κάθε δείγματος, γίνεται υπολογισμός ενός κανόνα απόφασης, ο οποίος αποτελεί είσοδο σε μία διαδικασία ελέγχου μεταξύ των υποθέσεων:

$$\Upsilon_0 : \theta = \theta_0$$

$$\Upsilon_1 : \theta = \theta_1$$

όπου το θ είναι η τιμή κεντρικού άξονα των δεδομένων, που στη συγκεκριμένη περίπτωση είναι η μετρική του μέσου όρου των δειγμάτων.

Εάν η διαδικασία απόφασης έχει ως αποτέλεσμα την υπόθεση Υ_0, συνεχίζεται η δειγματοληψία και ο έλεγχος. Εάν η διαδικασία απόφασης έχει ως αποτέλεσμα την υπόθεση Υ_1, η διαδικασία δειγματοληψίας διακόπτεται. Στην περίπτωση που το σύνολο δειγμάτων έχει σταθερό μέγεθος και είναι προκαθορισμένο, ο κανόνας απόφασης α_K δίδεται από:

$$\alpha_K = 0 \text{ εάν } A_1^N < \lambda, \text{επιλογή } \Upsilon_0$$

$$\alpha_K = 1 \text{ εάν } A_1^N \geq \lambda, \text{επιλογή } \Upsilon_1$$

όπου $K = 1, 2, 3, \dots, N$ και A_1^N είναι μία συνάρτηση απόφασης και λ είναι ένα κατώφλι, το οποίο έχει τιμή που διαφέρει και προσαρμόζεται ανάλογα με την εφαρμογή. Η απόφαση

λαμβάνεται με τη βοήθεια ενός κανόνα διακοπής, ο οποίος στην περίπτωση αυτή ορίζεται ως:

$$t_{συναγερμός} = N.\min\{K: \alpha_K = 1\}$$

όπου α_K είναι ο κανόνας απόφασης για τον αριθμό των δειγμάτων από το σύνολο δεδομένων K ή το προκαθορισμένο αριθμό δειγμάτων N και $t_{συναγερμός}$ είναι ο τρέχων αριθμός εκτέλεσης της διαδικασίας, στον οποίο παρουσιάστηκε συναγερμός απότομης μεταβολής.

Στην περίπτωση που η κατανομή των τιμών είναι Γκαουσιανή με μέση τιμή $\theta = \mu$ και σταθερή τιμή τυπικής απόκλισης σ^2, η συνάρτηση πυκνότητας πιθανότητας ορίζεται ως:

$$P_\theta(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

Γίνεται δημιουργία ενός διαγράμματος ελέγχου αλγορίθμου Shewhart, σύμφωνα με το οποίο όταν ισχύει $\mu_1 > \mu_0$, ο συναγερμός απότομης μεταβολής εμφανίζεται τη χρονική στιγμή που περιγράφεται από:

$$\bar{y}(K) \geq \mu_0 + \delta\frac{\sigma}{\sqrt{N}}, K = 1, 2, 3, \dots, N$$

όπου η τιμή του μέσου όρου των τιμών δειγμάτων μέχρι τη χρονική στιγμή κ είναι:

$$\bar{y}(K) = \frac{1}{N}\sum_{i=N(K-1)+1}^{NK} y_i$$

όπου δ και N είναι οι παραμέτροι ρύθμισης του διαγράμματος ελέγχου αλγορίθμου Shewhart, δηλαδή οι παράμετροι οι οποίοι ρυθμίζονται σύμφωνα με τα χαρακτηριστικά της εφαρμογής. Οι παράμετροι αυτοί μπορεί να οριστούν από την αρχή, αλλά συνήθως είναι πιο χρήσιμο να γίνει μία προσέγγιση τους με τη βοήθεια κάποιας διαδικασίας προεπεξεργασίας των δεδομένων.

Το μέγεθος $\mu_0 + \delta\frac{\sigma}{\sqrt{N}}$ είναι το ανώτατο όριο ελέγχου, το οποίο χρησιμοποιείται για ανίχνευση απότομης μεταβολής, η οποία προκύπτει από την αύξηση των τιμών δειγμάτων. Παρόλο που σε κάποιες εφαρμογές γίνεται χρήση μόνο του ανώτατου ορίου ελέγχου, στις περισσότερες των εφαρμογών είναι χρήσιμο να γίνεται ανίχνευση απότομης μεταβολής, η οποία προκύπτει από την αύξηση αλλά και τη μείωση των τιμών δειγμάτων. Στην περίπτωση αυτή, η τιμή του μέσου όρου μετά την απότομη μεταβολή μπορεί να είναι $\mu_1^+ = \mu_0 + v$ ή $\mu_1^- = \mu_0 - v$. Σε αυτή την περίπτωση, το κατώτατο όριο ελέγχου ορίζεται

ως $\mu_0 - \delta \frac{\sigma}{\sqrt{N}}$ και ο συναγερμός απότομης μεταβολής εμφανίζεται τη χρονική στιγμή που περιγράφεται από:

$$|\bar{y}(K) - \mu_0| \geq \delta \frac{\sigma}{\sqrt{N}}, K = 1, 2, 3, \dots, N$$

2.4.2 Ο αλγόριθμος διαγραμμάτων ελέγχου Shewhart

Τα διαγράμματα ελέγχου αλγορίθμου Shewhart [25] παρέχουν ένα στατιστικό μέτρο για την ανίχνευση απότομων μετατοπίσεων μίας χρονοσειράς. Στο διάγραμμα ελέγχου αλγορίθμου Shewhart, μία μεταβλητή x_t ανιχνεύεται να αποκλίνει σε χρόνο t από την κανονικότητα της, κάθε φορά που ξεπερνάει ένα από τα όρια ελέγχου, τα οποία καθορίζονται από τον αλγόριθμο: το ανώτατο όριο ελέγχου και το κατώτατο όριο ελέγχου.

Τα όρια ελέγχου ορίζονται ως η μετρικές απόστασης από την τρέχουσα τιμή μέσου όρου της στατιστικής διαδικασίας x_t. Συγκεκριμένα, το ανώτατο όριο ελέγχου και κατώτατο όριο ελέγχου ορίζονται ως:

$$AOE = \bar{x_t} + \kappa . \sigma_t$$

$$KOE = \bar{x_t} - \kappa . \sigma_t$$

όπου $\bar{x_t}$ αντιπροσωπεύει το μέσο όρο της χρονοσειράς σε χρόνο t και σ_t είναι η τυπική απόκλιση στο ίδιο χρονικό βήμα. Η παράμετρος κ αντιπροσωπεύει τη στεγανότητα της διαδικασίας ανίχνευσης μεταβολών. Χαμηλές τιμές κ οδηγούν σε στεγανό έλεγχο της διαδικασίας μετρήσεων, ενώ μεγάλες τιμές σηματοδοτούν μόνο τις μετρήσεις, οι οποίες είναι σε μεγάλο βαθμό εκτός ελέγχου.

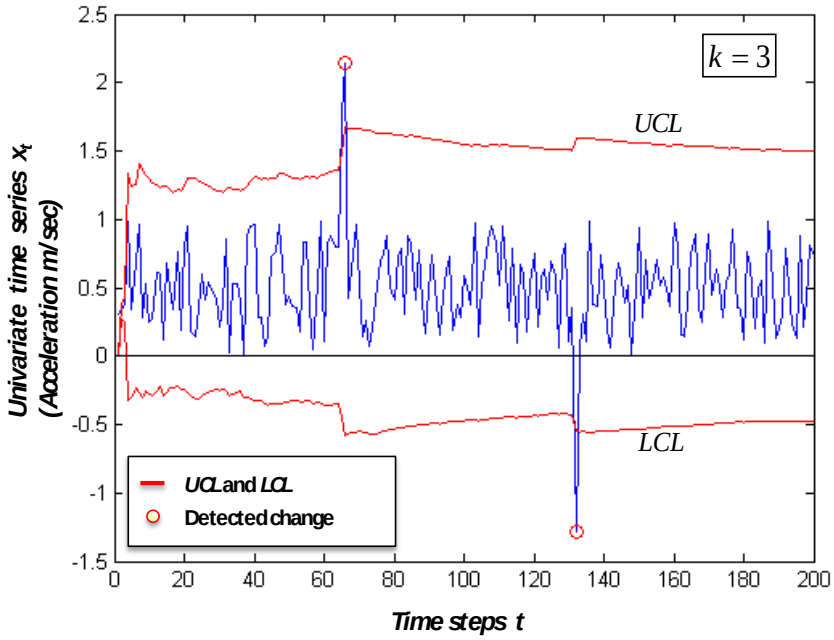
Σε χρονικό βήμα t η ροή δεδομένων x_t ανιχνεύεται να ενεργοποιεί ένα συναγερμό, εάν ισχύει $x_t > AOE$ ή $x_t < KOE$. Ο αλγόριθμος επιστρέφει ένα σήμα ανίχνευσης εξόδου σ σε κάθε βήμα t όπου $\sigma = 1$ εάν υπάρχει ανίχνευση μεταβολής. Στην περίπτωση φυσιολογικής συμπεριφοράς, δηλαδή $x_i \in [AOE, KOE]$, ο αλγόριθμος ορίζει $\sigma = 1$. Στη συνέχεια παρουσιάζεται ο αλγόριθμος διαγραμμάτων ελέγχου Shewhart.

Είσοδος: μονοδιάστατη χρονοσειρά x_t, στεγανότητα κ
΄Εξοδος: σήμα ανίχνευσης σ

$\bar{x}_0 \leftarrow 0$;
$\sigma_0 \leftarrow 0$;
$t \leftarrow 1$;
while (*true*)
    $\bar{x}_t \leftarrow \bar{x}_{t-1} + \frac{x_t - \bar{x}_{t-1}}{t}$;
    $\sigma_t \leftarrow \sqrt{\frac{1}{t}\left((t-1)\cdot\sigma_{t-1}^2 + \left(x_t - \bar{x}_t\right)\left(x_t - \bar{x}_{t-1}\right)\right)}$;
    $AOE_i \leftarrow \bar{x}_t + \kappa.\sigma_t$;
    $KOE_i \leftarrow \bar{x}_t - \kappa.\sigma_t$;
    if (($x_t > AOE$) **or** ($x_t < KOE$)) **then**
        $\sigma \leftarrow 1$;
    else
        $\sigma \leftarrow 0$;
    end
    $t \leftarrow t+1$;
end

Αλγόριθμος 2: Αλγόριθμος διαγραμμάτων ελέγχου Shewhart

Η Εικόνα 5 απεικονίζει ένα παράδειγμα του αλγορίθμου διαγραμμάτων ελέγχου Shewhart πάνω σε μία ροή αισθητήρων επιταχυνσιόμετρου Arduino. Με παρόμοιο τρόπο, όπως στον αλγόριθμο συσσωρευτικού αθροίσματος, στην περίπτωση χρονοσειρών πολλαπλών μεταβλητών, το διάγραμμα ελέγχου αλγορίθμου Shewhart πρέπει να εφαρμοστεί σε κάθε μεταβλητή ξεχωριστά. Ωστόσο, ο αλγόριθμος υποθέτει κανονική κατανομή της μεταβλητής x_t. Το γεγονός αυτό κάνει τον αλγόριθμο αρκετά εύρωστο σε σύνολα δεδομένων πραγματικού χρόνου, όπου στις περισσότερες των περιπτώσεων δεν υπάρχει κάποια διαθέσιμη πληροφορία για την κατανομή πιθανοτήτων την οποία ακολουθεί μία ροή δεδομένων. Από την άλλη πλευρά, ο αλγόριθμος είναι λιγότερο προσαρμοστικός σε σύγκριση με τον αλγόριθμο συσσωρευτικού αθροίσματος, αφού τα όρια ελέγχου μπορούν να τροποποιηθούν ελάχιστα στην περίπτωση μεγάλων χρονοσειρών.



Εικόνα 5: Αυθεντική ροή αισθητήρων καταμέτρησης επιτάχυνσης μέσω MPU και ανίχνευση μεταβολών με τον αλγόριθμο διαγραμμάτων ελέγχου Shewhart

2.5 Σύγκριση αλγορίθμου συσσωρευτικού αθροίσματος και αλγορίθμου διαγραμμάτων ελέγχου Shewhart

Λόγω της ποικιλομορφίας μαθηματικών πολυπλοκοτήτων των αλγορίθμων ανίχνευσης μεταβολών, αλλά και του μεγάλου αριθμού πιθανών εφαρμογών, η επιλογή των κατάλληλων αλγορίθμων γίνεται ανάλογα με τα χαρακτηριστικά της κάθε εφαρμογής. Ο αλγόριθμος διαγραμμάτων ελέγχου Shewhart μπορεί να ανιχνεύσει αποκλίσεις από την τιμή κεντρικού άξονα όταν αυτές είναι αρκετά μεγάλες, και σε χρονικές στιγμές οι οποίες είναι μεταγενέστερες των αποκλίσεων. Ο αλγόριθμος διαγραμμάτων ελέγχου Shewhart, στις περισσότερες των περιπτώσεων, δεν μπορεί να ανιχνεύσει αποκλίσεις από την τιμή κεντρικού άξονα όταν αυτές είναι μικρού μεγέθους. Μάλιστα, έχει παρατηρηθεί ότι η μη ανίχνευση των αποκλίσεων αυτών συμβαίνει ακόμα και στις περιπτώσεις που οι αποκλίσεις αυτές είναι επίμονες [37].

Ο αλγόριθμος συσσωρευτικού αθροίσματος είναι ευαίσθητος σε αποκλίσεις μετρίου ή μικρού μεγέθους από την τιμή κεντρικού άξονα, οι οποίες είναι επίμονες. Η τιμή κεντρικού άξονα μπορεί να βασίζεται στο μέσο όρο, στην τυπική απόκλιση και άλλα. Επίσης, η ανίχνευση των αποκλίσεων αυτών γίνεται σε χρονικές στιγμές οι οποίες δεν είναι πολύ μεταγενέστερες των αποκλίσεων. Με άλλα λόγια, η καθυστέρηση ανίχνευσης αποκλίσεων είναι σχετικά μικρή.

Σε σύγκριση με τον αλγόριθμο διαγραμμάτων ελέγχου Shewhart, ο αλγόριθμος συσσωρευτικού αθροίσματος ανιχνεύει μεταβολές με υψηλότερες πιθανότητες, όταν αυτές έχουν μικρή μεταβλητότητα, αλλά επιμονή στο χρόνο. Αυτό δικαιολογείται από το γεγονός ότι ο αλγόριθμος συσσωρευτικού αθροίσματος που περιεγράφηκε προηγουμένως συσσωρεύει λίγες επιδράσεις σε συνάρτηση με το χρόνο. Στο άρθρο [22] έγινε μία μελέτη του αλγόριθμου συσσωρευτικού αθροίσματος μη παραμετρικού χαρακτήρα, για μία συγκεκριμένη οικογένεια εκθετικών κατανομών. Υπό αυτές τις προϋποθέσεις, το χρονικό διάστημα καθυστέρησης ανίχνευσης μεταβολής προσεγγίζει το θεωρητικό ελάχιστο, εάν η τιμή μέσου χρόνου μεταξύ ψευδών συναγερμών ανίχνευσης τείνει στο άπειρο.

2.6 Αυτοπαλίνδρομο μοντέλο πολλαπλών μεταβλητών

Ένας περιορισμός, ο οποίος εμφανίζεται στους αλγορίθμους που έχουν παρουσιαστεί μέχρι τώρα, είναι η επιβολή μίας μονόδρομης σχέσης, ότι η προς διερεύνηση μέτρηση ή μεταβλητή επηρεάζεται από τις εσωτερικές μεταβλητές του αλγορίθμου, αλλά όχι από τις μεταβλητές των αλγορίθμων άλλων μετρήσεων. Ωστόσο, σε ένα μεγάλο αριθμό εφαρμογών, είναι χρήσιμο να υπάρχει μία συσχέτιση ή επηρεασμός μεταξύ των διαφόρων μετρήσεων. Για παράδειγμα, στην περίπτωση δύο μετρήσεων, μία αμφίδρομη σχέση επηρεασμού μεταξύ τους ίσως να είναι κατάλληλη, όπως για παράδειγμα η μέτρηση θερμοκρασίας και η μέτρηση υγρασίας στην ατμόσφαιρα. Τέτοιες σχέσεις επηρεασμού είναι επιτρεπτές στο πλαίσιο της παλινδρόμησης πολλαπλών μεταβλητών. Σε ένα τέτοιο πλαίσιο, όλες οι μεταβλητές αντιμετωπίζονται συμμετρικά. Η μοντελοποίηση των μεταβλητών είναι τέτοια, ώστε να υπάρχει η υπόθεση ότι κάθε μεταβλητή επηρεάζει κάθε άλλη μεταβλητή ισοδύναμα. Με μαθηματική ορολογία, οι μεταβλητές θεωρούνται ενδογενείς. Σε αυτό το πλαίσιο, οι μετρήσεις ή μεταβλητές ορίζονται ως ys: η μεταβλητή $y_{1,t}$ είναι η τιμή της μεταβλητής y_1 τη χρονική στιγμή t, η μεταβλητή $y_{2,t}$ είναι η τιμή της μεταβλητής y_2 τη χρονική στιγμή t.

Ένα αυτοπαλίνδρομο μοντέλο πολλαπλών μεταβλητών [26] είναι μία γενίκευση του μονομεταβλητού αυτοπαλίνδρομου μοντέλου για την πρόβλεψη ενός συνόλου μετρήσεων. Το σύνολο μετρήσεων μπορεί επίσης να εκφραστεί και ως ένα διάνυσμα στο χρόνο. Αποτελείται από μία εξίσωση για κάθε μεταβλητή του συστήματος, όπου κάθε εξίσωση αποτελείται από μία σταθερά και από προηγούμενα βήματα όλων των μεταβλητών του συστήματος. Εάν λάβουμε υπόψη ένα σύστημα παλινδρόμησης

πολλαπλών μεταβλητών με δύο μεταβλητές και ένα προηγούμενο βήμα, το σύστημα αυτό περιγράφεται ως εξής:

$$y_{1,t} = \delta_1 + \theta_{11,1} y_{1,t-1} + \theta_{12,1} y_{2,t-1} + \varepsilon_{1,t}$$

$$y_{2,t} = \delta_2 + \theta_{21,1} y_{1,t-1} + \theta_{22,1} y_{2,t-1} + \varepsilon_{2,t}$$

όπου $\varepsilon_{1,t}$ και $\varepsilon_{2,t}$ είναι διαδικασίες λευκού θορύβου, οι οποίες μπορεί να συσχετίζονται ταυτοχρόνως. Ο συντελεστής $\theta_{ii,\lambda}$ εκφράζει την επίδραση που έχει η μεταβλητή y_i στον εαυτό της σε λ προηγούμενα βήματα, ενώ ο συντελεστής $\theta_{ij,\lambda}$εκφράζει την επίδραση που έχει η μεταβλητή y_j στη μεταβλητή y_j σε λ προηγούμενα βήματα.

2.6.1 Παράμετροι αυτοπαλίνδρομου μοντέλου πολλαπλών μεταβλητών

Όταν το αυτοπαλίνδρομο μοντέλο πολλαπλών μεταβλητών χρησιμοποιείται για πρόβλεψη, υπάρχουν κάποιες παράμετροι, οι οποίες πρέπει να καθοριστούν [38]. Αυτές είναι, πόσες μεταβλητές κ και πόσα προηγούμενα βήματα β πρέπει να περιλαμβάνει το σύστημα. Ο αριθμός των παραμέτρων, οι οποίες πρέπει να χρησιμοποιηθούν και να γίνει μία εκτίμηση των τιμών τους σε ένα αυτοπαλίνδρομο μοντέλο πολλαπλών μεταβλητών ισούται με $\kappa + \beta\kappa^2$ ή $1 + \beta\kappa$ για κάθε εξίσωση. Εάν για παράδειγμα υπάρχουν 4 μεταβλητές προς διερεύνηση, και γίνεται μελέτη για τα προηγούμενα 2 βήματα, υπάρχουν 8 παράμετροι για κάθε εξίσωση, η τιμή των οποίων πρέπει να προσεγγιστεί, συνεπώς ο συνολικός αριθμός παραμέτρων που πρέπει να εκτιμηθούν είναι 32. Όσο μεγαλώνει ο αριθμός των παραμέτρων ή συντελεστών οι οποίοι πρέπει να προβλεφθούν, τόσο μεγαλώνει το σφάλμα εκτίμησης, το οποίο εισέρχεται στην πρόβλεψη.

Μία τυπική πρακτική στο αυτοπαλίνδρομο μοντέλο πολλαπλών μεταβλητών, έτσι ώστε το σφάλμα προσέγγισης από την πρόβλεψη των συντελεστών των εξισώσεων να είναι σε αποδεκτό βαθμό, είναι η τήρηση ενός μικρού αριθμού από μεταβλητές κ. Αυτό μπορεί να γίνει δυνατό εάν στο μοντέλο συμπεριληφθούν μόνο μεταβλητές, οι οποίες συσχετίζονται μεταξύ τους, και συνεπώς χρήσιμες για την πρόβλεψη μεταξύ τους. Όσο αφορά τον αριθμό των προηγούμενων βημάτων που πρέπει να χρησιμοποιηθούν για την πρόβλεψη, μία τυπική πρακτική είναι η χρήση κριτηρίων, τα οποία σχετίζονται με τη διαθέσιμη πληροφορία.

2.6.2 Ο αλγόριθμος αυτοπαλίνδρομου μοντέλου πολλαπλών μεταβλητών

Εάν ληφθεί υπόψη η υπόθεση μίας χρονοσειράς μίας μεταβλητής x_t, οι διαδοχικές μετρήσεις τέτοιων σειρών περιέχουν πληροφορία, η οποία είναι σχετική με τη διαδικασία που τις δημιούργησε. Μία προσπάθεια για την περιγραφή της υποκείμενης αυτής διάταξης μπορεί να επιτευχθεί με τη μοντελοποίηση της τρέχουσας τιμής της μεταβλητής x_t ως ένα σταθμισμένο γραμμικό άθροισμα των προηγούμενων τιμών της, όπως για παράδειγμα x_{t-1} ή x_{t-2}. Αυτή είναι μία αυτοπαλίνδρομη διαδικασία, και αποτελεί μία πολύ απλή, αλλά αποτελεσματική προσέγγιση για το χαρακτηρισμό χρονοσειρών [26].

Η τάξη του μοντέλου είναι ο αριθμός των προηγούμενων παρατηρήσεων, οι οποίες χρησιμοποιούνται για τον προσδιορισμό του x_t. Τα βάρη είναι οι παράμετροι του μοντέλου, των οποίων οι τιμές εκτιμήθηκαν από τα δεδομένα που χαρακτηρίζουν με μοναδικό τρόπο τη χρονοσειρά. Τα αυτοπαλιδρόμενα μοντέλα πολλαπλών μεταβλητών επεκτείνουν αυτή την προσέγγιση σε πολλαπλές χρονοσειρές, σε σημείο που το διάνυσμα $x_t = (x_{1,t}, x_{2,t}, \dots, x_{n,t})$ των τρέχοντων τιμών όλων των μεταβλητών μοντελοποιείται ως ένα γραμμικό άθροισμα των προηγούμενων διανυσμάτων. Το αυτοπαλίνδρομο μοντέλο πολλαπλών μεταβλητών περιγράφει τη δυναμική συμπεριφορά της χρονοσειράς για πρόβλεψη [26].

Εάν το $x_t = (x_{1,t}, x_{2,t}, \dots, x_{n,t})$ δηλώνει ένα n διαστάσεων διάνυσμα πλαισίου αποτελούμενο από στοιχεία πλαισίου χρονοσειράς $x_{i,t}$ όπου $i = 1, \dots, n$ και $T = 1,2, \dots$ είναι το σύνολο από διακριτά χρονικά βήματα. Σύμφωνα με ένα τάξης θ μοντέλο παλινδρόμησης πολλαπλών μεταβλητών $ΜΠΠΜ(\theta)$, κάθε διάνυσμα πλαισίου x_t αντιπροσωπεύεται από τη μορφή $x_t = \Pi_\sigma + \Pi_1 x_{t-1} + \cdots + \Pi_\theta x_{t-\theta} + \varepsilon_t$. Οι μεταβλητές Π_i όπου $i \in [1, \varphi]$ είναι οι $(n \times n)$ συντελεστές πίνακα, Π_σείναι ένα σταθερό διάνυσμα και ε_t είναι μία $(n \times 1)$ μη παρατηρήσιμη, μηδενικού μέσου όρου διαδικασία λευκού θορύβου, η οποία είναι σειριακά ανεξάρτητη και με πίνακα συνδιακύμανσης ο οποίος είναι αμετάβλητος στο χρόνο.

Ένας από τους κύριους στόχους των αυτοπαλίνδρομων μοντέλων πολλαπλών μεταβλητών είναι η πρόβλεψη. Εάν ληφθεί υπόψη η υπόθεση ότι μπορεί να γίνει μία καλή προσέγγιση των παραμέτρων της $ΜΠΠΜ(\theta)$ διαδικασίας [39], η καλύτερη γραμμική διαδικασία πρόβλεψης, όσο αφορά το ελάχιστο μέσο τετραγωνικό σφάλμα του διανύσματος πλαισίου x_t, για παράδειγμα πρόβλεψη ενός βήματος, με βάση τη διαθέσιμη πληροφορία σε χρόνους μέχρι και $t - \theta$ μπορούν να υπολογιστούν ως εξής:

$$\widetilde{x_t} = \Pi_\sigma + \Pi_1 x_{t-1} + \cdots + \Pi_\theta x_{t-\theta}$$

Για κάθε εκτιμώμενη μεταβλητή $\widetilde{x_i}$ του προβλεπόμενου διανύσματος $\widetilde{x_t}$, ορίζεται ως σχετικό σφάλμα πρόβλεψης $e_{i,t}$ το σφάλμα που λαμβάνεται από την σύγκριση της μεταβλητής $\widetilde{x_i}$ με την αντίστοιχη μεταβλητή x_i του πραγματικού διανύσματος x_t. Το σχετικό σφάλμα πρόβλεψης $e_{i,t}$ μπορεί να υπολογιστεί όπως στη συνέχεια:

$$e_{i,t} = \frac{\left\| x_{i,t} - \widetilde{x_{i,t}} \right\|}{\left\| x_{i,t} \right\|}$$

όπου $\|x\|$ είναι η Ευκλείδια κανονική μορφή της μεταβλητής x. Το σχετικό σφάλμα είναι ευαίσθητο σε χαμηλές τιμές των πραγματικών ροών δεδομένων, δεδομένου ότι σε περίπτωση μίας τιμής η οποία είναι κοντά στη μηδενική, το σχετικό σφάλμα θα μπορούσε να αυξηθεί ραγδαία ακόμα και για μικρές τιμές απόλυτου σφάλματος. Σε κάθε βήμα, μετά τον υπολογισμό του σχετικού σφάλματος πρόβλεψης για κάθε x_i, πρέπει να γίνει σύγκριση της τιμής αυτής με μία προκαθορισμένη τιμή κατωφλίου $κατώφλι_i$, προκειμένου να προσδιοριστεί εάν υπάρχει σημαντική απόκλιση μεταξύ των μεταβλητών του εκτιμώμενου διανύσματος πλαισίου και των πραγματικών που καταφθάνουν, απόκλιση η οποία υποδηλώνει κάποια μεταβολή. Ο αλγόριθμος ανίχνευσης μεταβολών αυτοπαλίνδρομου μοντέλου πολλαπλών μεταβλητών παρουσιάζεται στη συνέχεια:

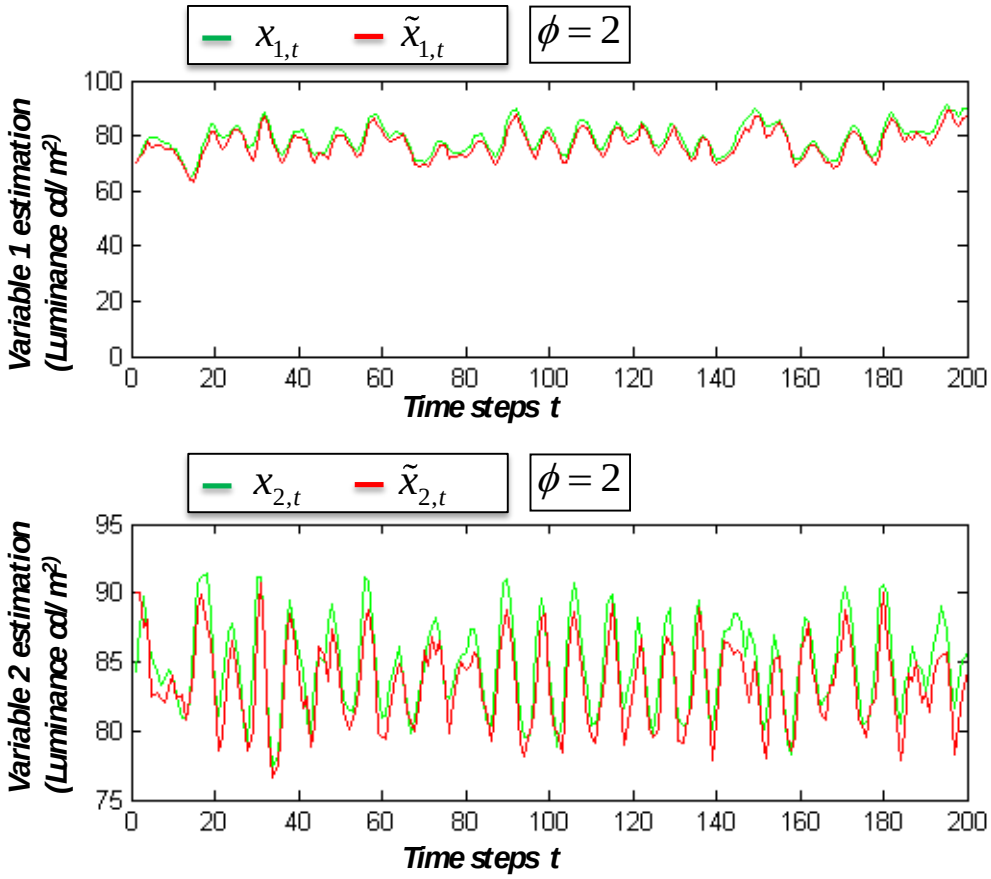
Είσοδος: χρονοσειρά πολλαπλών μεταβλητών $x_t = (x_{1,t}, x_{2,t}, \ldots, x_{n,t})$, αριθμός δειγμάτων εκπαίδευσης k, τιμές κατωφλίου $κατώφλι_1, \ldots, κατώφλι_n$

Έξοδος: σήμα ανίχνευσης σ

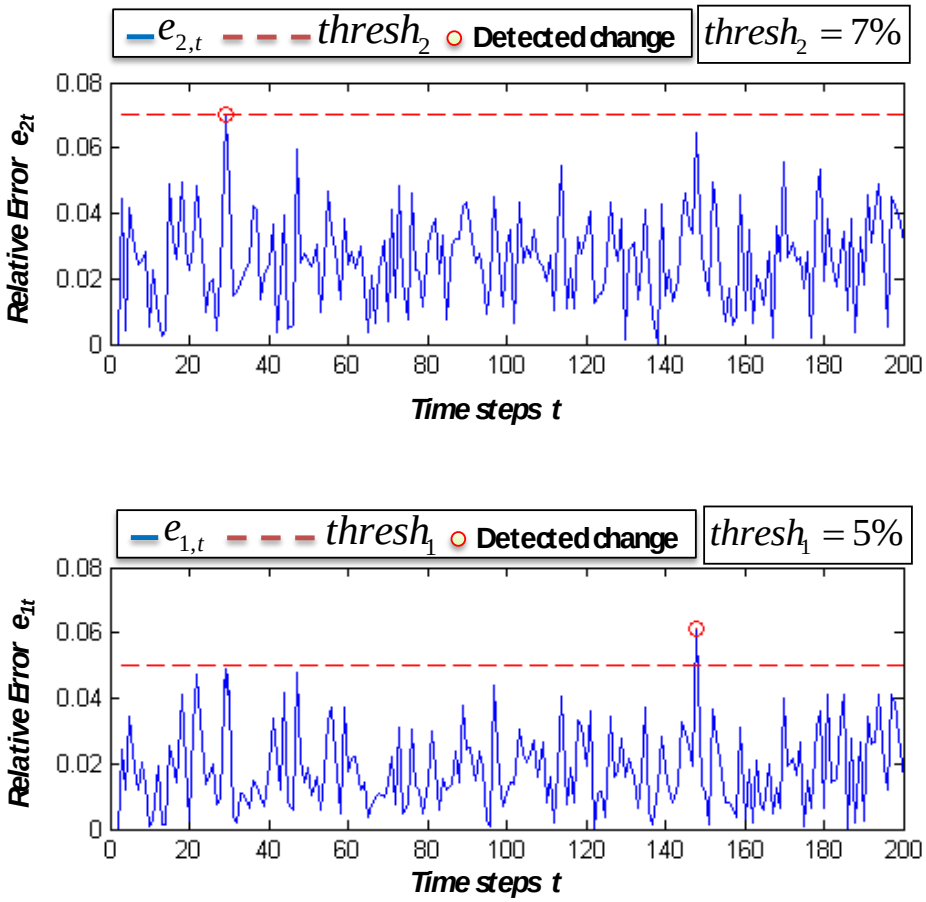
Προσέγγιση του μοντέλο σύμφωνα με τα δεδομένα εκπαίδευσης $\{\mathbf{x}_t\},\ \forall t \in [1,k]$
$t \leftarrow k+1$;
while (*true*)
    $x_t = \Pi_\sigma + \Pi_1 x_{t-1} + \cdots + \Pi_\theta x_{t-\theta}$;
    for $i \leftarrow 1$ **to** n
        $e_{i,t} \leftarrow \frac{\left\| x_{i,t} - \tilde{x}_{i,t} \right\|}{\left\| x_{i,t} \right\|}$;
        if ($e_{i,t} > κατώφλι_i$) **then**
            $\sigma \leftarrow 1$;
        else
            $\sigma \leftarrow 0$;
        end
    end
end

Αλγόριθμος 3: Αλγόριθμος ανίχνευσης μεταβολών αυτοπαλινδρόμενου μοντέλου πολλαπλών μεταβλητών

Οι Εικόνες 6 και 7 απεικονίζουν ένα παράδειγμα της προσέγγισης ανίχνευσης μεταβολών που βασίζεται στο αυτοπαλίνδρομο μοντέλο πολλαπλών μεταβλητών. Σε αυτό το παράδειγμα, γίνεται υπόθεση ενός δισδιάστατου διανύσματος πλαισίου $x_t = (x_{1,t}, x_{2,t})$ όπου σε κάθε χρονικό βήμα, το εκτιμώμενο διάνυσμα πλαισίου $\widetilde{x_t} = (\widetilde{x_{1,t}}, \widetilde{x_{2,t}})$ υπολογίζεται μέσω ενός τάξης δύο αυτοπαλίνδρόμενου μοντέλου πολλαπλών μεταβλητών. Οι πράσινες γραμμές αντιπροσωπεύουν τις πραγματικές ροές δεδομένων μίας μεταβλητής $x_{1,t}, x_{2,t}$, οι οποίες καταφθάνουν σε κάθε χρονικό βήμα t. Οι κόκκινες γραμμές απεικονίζουν τις εκτιμώμενες ροές δεδομένων μίας μεταβλητής $\widetilde{x_{1,t}}$, $\widetilde{x_{2,t}}$. Στην Εικόνα 7 γίνεται ανίχνευση μεταβολής σε κάθε μεταβλητή του διανύσματος πλαισίου, κάθε φορά που το σχετικό σφάλμα πρόβλεψης υπερβαίνει ένα κατώφλι. Σε αυτό το παράδειγμα, τα σχετικά σφάλματα διαφέρουν για τις δύο μεταβλητές, από τις οποίες αποτελείται το διάνυσμα πλαισίου.



Εικόνα 6: Χρονοσειρά δύο διαστάσεων η οποία αναπαριστά τιμές φωτεινότητας και εκτίμηση με βάση ένα δεύτερης τάξης αυτοπαλινδρόμενο μοντέλο πολλαπλών μεταβλητών



Εικόνα 7: Ανίχνευση μεταβολών με βάση το αυτοπαλίνδρομο μοντέλο πολλαπλών μεταβλητών σε χρονοσειρά δύο διαστάσεων

2.6.3 Χαρακτηριστικά αυτοπαλίνδρομου μοντέλου πολλαπλών μεταβλητών

Ένα σημαντικό μειονέκτημα το οποίο παρουσιάζεται στη βιβλιογραφία όσον αφορά τα μοντέλα παλινδρόμησης πολλαπλών μεταβλητών είναι ότι δε χρησιμοποιούν κάποιου είδους θεωρία ή επιπρόσθετη πληροφορία για την πρόβλεψη [40]. Η παλινδρόμηση πολλαπλών μεταβλητών δεν κατασκευάζεται με βάση κάποια θεωρία, όπως για παράδειγμα θεωρία από την επιστήμη της βιολογίας ή φυσικής, η οποία μπορεί να επιβάλει κάποια θεωρητική δομή στις επιμέρους εξισώσεις. Ακόμα ένα σημαντικό μειονέκτημα που παρουσιάζεται είναι η υπόθεση ότι κάθε μεταβλητή επηρεάζει κάθε άλλη μεταβλητή στο σύστημα, κάτι που σε ένα μεγάλο αριθμό εφαρμογών δεν ισχύει. Η υπόθεση αυτή δυσχεραίνει την προσπάθεια εκτίμησης των συντελεστών ή παραμέτρων του συστήματος και μπορεί να οδηγήσει σε αύξηση του σφάλματος πρόβλεψης.

Το μοντέλο παλινδρόμησης πολλαπλών μεταβλητών παρουσιάζει κάποια μειονεκτήματα τα οποία μπορεί να το καθιστούν μη χρήσιμο για κάποιες εφαρμογές, ωστόσο για κάποιες κατηγορίες εφαρμογών το μοντέλο αυτό είναι χρήσιμο [38]. Για παράδειγμα, το μοντέλο μπορεί να χρησιμοποιηθεί για πρόβλεψη ενός συνόλου από συσχετιζόμενες μεταβλητές, για τις οποίες δεν απαιτείται κάποια ρητή ερμηνεία. Επίσης, μπορεί να χρησιμοποιηθεί στην ανάλυση της διακύμανσης του σφάλματος πρόβλεψης, όπου ένα ποσοστό της διακύμανσης πρόβλεψης της μία μεταβλητής αποδίδεται στην επίδραση των άλλων μεταβλητών. Η ανάλυση αιφνίδιων αντιδράσεων, στην οποία αναλύεται η αντίδραση μίας μεταβλητής σε μία ξαφνική, αλλά προσωρινή, αλλαγή κάποιας άλλης μεταβλητής είναι ακόμα μία εφαρμογή, στην οποία το μοντέλο παλινδρόμησης πολλαπλών μεταβλητών μπορεί να εφαρμοστεί. Το μοντέλο είναι επίσης χρήσιμο για τον έλεγχο εάν μία μεταβλητή είναι χρήσιμη στην πρόβλεψη κάποιας άλλης μεταβλητής.

2.7 Αλγόριθμος Εκθετικά Σταθμισμένου Κινούμενου Μέσου Όρου

Ο αλγόριθμος διαγράμματος ελέγχου εκθετικά σταθμισμένου κινούμενου μέσου όρου, ο οποίος περιγράφηκε πρώτα στο [41], είναι μία πραγματικού χρόνου μέθοδος ανίχνευσης απότομων μεταβολών [42]. Εάν γίνει η υπόθεση ότι υπάρχουν παρατηρηθείσες τιμές $x_1, x_2, \ldots$ οι οποίες είναι αποτέλεσμα δειγματοληψίας μίας κατανομής με γνωστά τα μεγέθη μέσου όρου μ και τυπικής απόκλισης σ^2. Ορίζονται οι νέες μεταβλητές $y_1, y_2, \ldots$ ως ακολούθως:

$$y_o = \mu$$

$$y_\kappa = (1-\lambda)y_{\kappa-1} + \lambda x_\kappa$$

όπου το λ έχει το ρόλο παράγοντα εκθετικής λήθης και ισχύει $\lambda \in [0,1]$. Μπορεί να αποδειχθεί ότι η τυπική απόκλιση της y_κ ισούται με:

$$\sigma_{y_\kappa} = \left(\sqrt{\frac{\rho}{2-\rho}[1-(1-\rho)^{2\kappa}]}\right)\sigma$$

Εάν στόχος είναι η ανίχνευση αύξησης στο μέσο όρο, μία απότομη μεταβολή ανιχνεύεται όταν $y_\kappa > \mu + \varepsilon\sigma_{y_\kappa}$ όπου ε είναι μία παράμετρος ελέγχου, η οποία έχει τιμή τέτοια ώστε ο αλγόριθμος να έχει την επιθυμητή απόδοση όσο αφορά τις μετρικές MAE_0 και MAE_1. Υπάρχει το ενδεχόμενο μεταβολής του αλγορίθμου εκθετικά σταθμισμένου κινούμενου μέσου όρου, έτσι ώστε να εκτελεστεί αμφίπλευρη ανίχνευση. Σύμφωνα με τη

βιβλιογραφία, μία τυπική τιμή της παραμέτρου ρ βρίσκεται στο διάστημα $(0{,}05\,,0{,}5)$ για ανίχνευση μικρών μετατοπίσεων, ενώ η παράμετρος ε έχει τυπικά μία τιμή, η οποία είναι κοντά στο τρία. Πρακτικά, ο αλγόριθμος είναι ιδανικός για ανίχνευση μικρών μετατοπίσεων στο μέσο όρο της διαδικασίας. Τα χαρακτηριστικά συμπεριφοράς του αλγορίθμου είναι παρόμοια με αυτά του αλγορίθμου συσσωρευτικού αθροίσματος. Ένα βασικό πρόβλημα, το οποίο εμφανίζεται και στους δύο αλγόριθμους είναι η επιλογή παραμέτρων ελέγχου.

3. ΣΥΣΧΕΤΙΣΗ ΣΥΜΒΑΝΤΩΝ ΣΕ ΡΟΕΣ ΔΕΔΟΜΕΝΩΝ ΣΥΜΒΑΝΤΩΝ ΠΟΛΛΑΠΛΩΝ ΜΕΤΑΒΛΗΤΩΝ

Κατά τις τελευταίες δεκαετίες, αρκετή μελέτη έχει αφιερωθεί στη συσχέτιση δεδομένων συμβάντων, τα οποία προέρχονται από υποδομές της τεχνολογίας της πληροφορίας, με στόχο την εξαγωγή προτύπων και την προληπτική δράση. Τα πρότυπα αυτά απεικονίζουν την εσωτερική δυναμική των προς διερεύνηση συστημάτων. Τυπικά συστήματα συσχέτισης συμβάντων, τα οποία λειτουργούν πάνω σε μονοδιάστατες χρονοσειρές, εκτελούνται κάτω από κάποιες προυποθέσεις. Η μετάβαση από ένα αντικείμενο X, το οποίο μπορεί να είναι ένα συμβάν ή μία ακολουθία από συμβάντα, σε ένα άλλο αντικείμενο Ψ, εμφανίζεται εάν και μόνο εάν το αντικείμενο Ψ εμφανίζεται ακριβώς μετά από το αντικείμενο X, και όχι για παράδειγμα εντός ενός χρονικού παραθύρου. Επίσης, σε κάθε βήμα της ακολουθίας, λαμβάνεται υπόψη μόνο ένα αντικείμενο. Με άλλα λόγια, δεν υπάρχουν αντικείμενα, τα οποία εμφανίζονται την ίδια χρονική στιγμή.

Ωστόσο, τα πράγματα μπορεί να διαφέρουν στην περίπτωση συσχέτισης συμβάντων πάνω σε δεδομένα αισθητήρων πολλαπλών μεταβλητών. Δεν υπάρχει κάποια εγγύηση ότι τα συμβάντα εμφανίζονται ένα σε κάθε χρονική στιγμή. Αντιθέτως, μία κατάσταση συναγερμού ή ένα δυσλειτουργικό σύστημα αναμένεται να οδηγήσει σε εμφάνιση ενός αριθμού από συμβάντα κατά το ίδιο χρονικό βήμα. Για παράδειγμα, εάν γίνεται παρακολούθηση μιας πυρκαγιάς μέσω ενός αριθμού από ανιχνευτές καπνού και αισθητήρες θερμοκρασίας και υγρασίας, μπορεί να γίνει εμφάνιση περισσότερων από ένα συμβάντων στο ίδιο χρονικό βήμα. Αυτό οφείλεται στο γεγονός ότι, οι περισσότεροι των αισθητήρων που παρακολουθούν μια επηρεαζόμενη περιοχή αναμένεται να παρουσιάσουν μία σημαντική μετατόπιση στην κατανομή πιθανοτήτων των αναφερόμενων τιμών τους. Στη συνέχεια των παραπάνω, στη συσχέτιση συμβάντων πάνω σε δεδομένα αισθητήρων, πρέπει να ληφθεί υπόψη και η πιθανότητα να μην υπάρχει εμφάνιση συμβάντων σε κάποια χρονικά βήματα. Αυτό οφείλεται στο γεγονός ότι σπάνια εμφανίζονται συμβάντα, όταν δε συμβαίνει κάποια κατάσταση συναγερμού.

Σε αυτή την ενότητα θα γίνει προσπάθεια μελέτης των προαναφερθέντων προκλήσεων, στο πλαίσιο των δεδομένων ροής συμβάντων πολλαπλών μεταβλητών. Λαμβάνεται υπόψη η υπόθεση ότι η εμφάνιση ροών διανυσμάτων συμβάντων είναι μία στοχαστική διαδικασία, και γίνεται συζήτηση ενός συστήματος συσχέτισης συμβάντων. Το σύστημα αυτό εκτελεί ένα μεταβλητής τάξης αλγόριθμο, ο οποίος βασίζεται στην ιδέα

της μερικής αντιστοίχισης [43], η οποία έχει προσαρμοστεί στο πλαίσιο πολυμεταβλητών χρονοσειρών. Στις επόμενες παραγράφους γίνεται συζήτηση της προσέγγισης αυτής και επίσης διερεύνηση της χρονικής πολυπλοκότητας του αλγορίθμου ανά περίπτωση. Πριν την περιγραφή του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών συμβάντων, γίνεται μελέτη του θεωρητικού υποβάθρου πάνω στο οποίο βασίζεται ο αλγόριθμος.

3.1 Πρόβλεψη με μερική αντιστοίχιση

Η μέθοδος πρόβλεψης με μερική αντιστοίχιση (prediction by partial matching) [1] είναι μία προσαρμοστική μέθοδος συμπίεσης, η οποία βασίζεται στη δημιουργία κάποιου μοντέλου με σκοπό την πρόβλεψη. Τα μοντέλα πρόβλεψης με μερική αντιστοίχιση χρησιμοποιούν μία σειρά από χαρακτήρες, οι οποίοι εμφανίστηκαν σε παρελθοντικά βήματα, για να προβλέψουν το επόμενο χαρακτήρα στη σειρά. Η μέθοδος πρόβλεψης με μερική αντιστοίχιση μπορεί επίσης να χρησιμοποιηθεί σε εφαρμογές κατηγοριοποίησης για ταξινόμηση των δεδομένων σε ομάδες, οι οποίες διαμορφώνονται ανάλογα με τα δεδομένα.

Η κεντρική ιδέα της μεθόδου πρόβλεψης με μερική αντιστοίχιση είναι να παρέχει ένα σχετικά ακριβές στατιστικό μοντέλο για αριθμητική κωδικοποίηση, και συνεπώς να επιτυγχάνεται αποδοτική συμπίεση [44] [45]. Η μέθοδος ανήκει στην κλάση μεθόδων συμπίεσης κειμένου των στατιστικών κωδικοποιητών. Οι στατιστικοί κωδικοποιητές κωδικοποιούν κάθε χαρακτήρα ανεξάρτητα, λαμβάνοντας υπόψη το παρελθοντικό τους πλαίσιο, δηλαδή τους προηγούμενους χαρακτήρες. Στη συνέχεια, η μέθοδος δημιουργεί ένα στατιστικό μοντέλο πλαισίου, το οποίο χρησιμοποιείται για υπολογισμό των πιθανοτήτων που προκύπτουν. Όσο αυξάνεται ο αριθμός των χαρακτήρων που περιέχονται στο παρελθοντικό πλαίσιο, τόσο μειώνεται ο αριθμός των πιθανοτήτων που υπολογίζονται, και συνεπώς βελτιώνεται η συμπίεση. Οι μέθοδοι στατιστικών κωδικοποιητών έχουν καλύτερη αποδοτικότητα συμπίεσης από τους λεξιλογικούς κωδικοποιητές, οι οποίοι χρησιμοποιούν τη μέθοδο συρόμενου παραθύρου για συμπίεση, ωστόσο απαιτούν μεγάλη χρησιμοποίηση της μνήμης τυχαίας προσπέλασης.

Η μέθοδος πρόβλεψης με μερική αντιστοίχιση χρησιμοποιεί την τεχνική της μοντελοποίησης πεπερασμένου πλαισίου, η οποία είναι μία μέθοδος που αναθέτει σε κάθε χαρακτήρα μία πιθανότητα, η οποία προκύπτει από το πλαίσιο στο οποίο εμφανίζεται. Το πλαίσιο στην περίπτωση αυτή ορίζεται ως τα σύμβολα που εμφανίστηκαν

σε παρελθοντικά βήματα του προς διερεύνηση χαρακτήρα. Η τάξη του μοντέλου ορίζεται ίση με το μέγεθος πλαισίου του μοντέλου, με άλλα λόγια είναι ίση με τον αριθμό παρελθοντικών βημάτων που λαμβάνονται υπόψη. Για παράδειγμα, όταν πρέπει να κωδικοποιηθεί το γράμμα η από τη λέξη $\mu\acute{\varepsilon}\tau\rho\eta\sigma\eta$ το μοντέλο πλαισίου τάξεως 4 δίνεται ως $\tau\rho\eta\sigma$, το μοντέλο πλαισίου τάξης 3 δίνεται ως $\rho\eta\sigma$ και ούτω κάθε εξής. Το μοντέλο πλαισίου επιτρέπει την πρόβλεψη χαρακτήρων με υψηλότερη πιθανότητα, συνεπώς απαιτείται μικρότερος χώρος μνήμης για κωδικοποίηση ενός χαρακτήρα. Παρόλο που η μέθοδος πρόβλεψης με μερική αντιστοίχιση δημιουργήθηκε για συμπίεση, μπορεί να είναι επίσης χρήσιμη και στην περίπτωση πρόβλεψης δεδομένων αισθητήρων.

Η μέθοδος πρόβλεψης με μερική αντιστοίχιση αποτελείται από δύο συστατικά, ένα στατιστικό μοντέλο πλαισίου και ένα αριθμητικό κωδικοποιητή [45]. Κάθε χαρακτήρας κωδικοποιείται ξεχωριστά, λαμβάνοντας υπόψη το πλαίσιο στο οποίο εμφανίζεται. Το μοντέλο πλαισίου αποθηκεύει τη συχνότητα εμφάνισης κάθε χαρακτήρα, και κάνει ταξινόμηση των χαρακτήρων σε κλαδιά δένδρων. Συνεπώς, κάθε χαρακτήρας σε ένα δένδρο πλαισίου έχει δύο μετρητές, ένα μετρητή για τους χαρακτήρες στα αριστερά $\mathit{ΜετρητήςΑριστερά}$ και ένα μετρητή για τους χαρακτήρες στα δεξιά $\mathit{ΜετρητήςΔεξιά}$. Για ένα χαρακτήρα, οι μετρητές αυτοί δίνονται από:

$$\mathit{ΜετρητήςΑριστερά}(i) = \sum_{\forall j<i} \mathit{καταμέτρηση}(\mathit{χαρακτήρας}(j))$$

$$\mathit{ΜετρητήςΔεξιά}(i) = \mathit{ΜετρητήςΑριστερά}(i) + \mathit{καταμέτρηση}(\mathit{χαρακτήρας}(i))$$

όπου $\mathit{καταμέτρηση}(\mathit{χαρακτήρας}(i))$ είναι η στατιστική καταμέτρηση του χαρακτήρα. Οι δύο στατιστικοί μετρητές είναι απαραίτητοι όταν γίνεται ερώτηση στο μοντέλο, με βάση ένα χαρακτήρα για τον οποίο υπάρχει πληροφορία πλαισίου.

Όταν ένας χαρακτήρας πρέπει να κωδικοποιηθεί, γίνεται έλεγχος του μοντέλου για το χαρακτήρα με πληροφορία πλαισίου της τάξης αυτού. Εάν ο χαρακτήρας βρίσκεται στο μοντέλο, γίνεται ανάκτηση των μετρητών $\mathit{ΜετρητήςΑριστερά}$ και $\mathit{ΜετρητήςΔεξιά}$ και γίνεται κωδικοποίηση του χαρακτήρα. Εάν ο χαρακτήρας δεν βρίσκεται στο μοντέλο, γίνεται μετάδοση ενός χαρακτήρα διαφυγής και γίνεται έλεγχος του μοντέλου με βάση τη επόμενη μικρότερη τάξη πλαισίου. Ο χαρακτήρας διαφυγής χρησιμοποιείται για να σηματοδοτήσει στον αποκωδικοποιητή την τρέχουσα τάξη πλαισίου του μοντέλου. Όπως κάθε χαρακτήρας έχει δύο μετρητές, έτσι και ο χαρακτήρας διαφυγής έχει $\mathit{ΜετρητήςΑριστερά}$ και $\mathit{ΜετρητήςΔεξιά}$. Με αυτό τον τρόπο, επιτυγχάνεται η

κωδικοποίηση του χαρακτήρα διαφυγής όπως γίνεται κωδικοποίηση όλων των υπόλοιπων χαρακτήρων. Οι μετρητές δίνονται από:

$$МετρητήςΑριστερά_{ΧΔ} = \sum_{\forall i} καταμέτρηση(χαρακτήρας(i))$$

$$МετρητήςΔεξιά_{ΧΔ} = МετρητήςΑριστερά_{ΧΔ} + διαφορετικός$$

Όπου $διαφορετικός$ δηλώνει τον αριθμό των διαφορετικών χαρακτήρων. Ο $МετρητήςΑριστερά$ ενός χαρακτήρα διαφυγής είναι το άθροισμα των δεξιών μετρήσεων των χαρακτήρων στο πλαίσιο. Στο $МετρητήςΔεξιά$ προστίθεται ο αριθμός των διαφορετικών χαρακτήρων στο πλαίσιο. Εάν ένας χαρακτήρας δεν βρίσκεται ούτε στο μοντέλο μηδενικής τάξης, γίνεται κωδικοποίηση μείον ένα τάξης, σύμφωνα με την οποία κάθε χαρακτήρας έχει ισοδύναμη πιθανότητα.

Το μοντέλο πλαισίου εφαρμόζεται προσαρμοστικά τόσο στο μηχανισμό κωδικοποίησης, όσο και στο μηχανισμό αποκωδικοποίησης. Όταν γίνεται κωδικοποίηση και αποκωδικοποίηση ενός χαρακτήρα, το μοντέλο βρίσκεται στην ίδια κατάσταση και στους δύο μηχανισμούς. Το μοντέλο μπορεί να υλοποιηθεί με διασυνδεδεμένους κόμβους μέσα σε ένα δένδρο δεδομένων. Όταν ένας χαρακτήρας με συγκεκριμένο πλαίσιο δεν ευρίσκεται στο δένδρο, αυτό μπορεί να οφείλεται είτε στη μη ύπαρξη του χαρακτήρα στο πλαίσιο, είτε στη μη ύπαρξη χαρακτήρων στο πλαίσιο, είτε στη μη ύπαρξη χαρακτήρων στο πλαίσιο και μη ύπαρξη του πλαισίου. Στην περίπτωση μη ύπαρξης του χαρακτήρα στο πλαίσιο, γίνεται δημιουργία ενός νέου κόμβου και αποστολή ενός χαρακτήρα διαφυγής. Στην περίπτωση μη ύπαρξης χαρακτήρων στο πλαίσιο, γίνεται δημιουργία ενός νέου κόμβου αλλά όχι αποστολή χαρακτήρα διαφυγής, αφού ο αποκωδικοποιητής μπορεί να συμπεράνει ότι πρέπει να κάνει μετάβαση σε χαμηλότερη τάξη χωρίς το χαρακτήρα διαφυγής. Στην περίπτωση μη ύπαρξης χαρακτήρων στο πλαίσιο και μη ύπαρξη πλαισίου, γίνεται δημιουργία αρχικά του πλαισίου και μετά νέου κόμβου.

Κάθε κόμβος του δένδρου αποθηκεύει ένα μετρητή του αριθμού εμφάνισης ενός χαρακτήρα. Όταν γίνεται εύρεση ενός χαρακτήρα, οι μετρητές $МετρητήςΑριστερά$. $МετρητήςΔεξιά$ και $МετρητήςΣύνολο$ πρέπει να υπολογίζονται. Αυτό μπορεί να γίνει με διάσχιση όλων των χαρακτήρων του πλαισίου. Εάν ευρεθεί ο χαρακτήρας που πρέπει να κωδικοποιηθεί, γίνεται αποθήκευση των μετρητών $МετρητήςΑριστερά$ και $МετρητήςΔεξιά$. Στη συνέχεια, γίνεται διάσχιση για τους υπόλοιπους χαρακτήρες του

πλαισίου, υπολογίζοντας ταυτοχρόνως το $Μετρητής Σύνολο$, που στη βιβλιογραφία επίσης αναφέρεται και ως συσσωρευτικός μετρητής. Όταν γίνεται διάσχιση των χαρακτήρων του πλαισίου, γίνεται ταυτόχρονη εφαρμογή της εξίσωσης $Μετρητής Αριστερά_{χΔ} = \sum_{\forall i} καταμέτρηση(χαρακτήρας(i))$. Για βελτίωση της αναζήτησης στο δένδρο, υπάρχουν διάφορες τεχνικές, όπως η διατήρηση επιπλέον δεικτών στους κόμβους, οι οποίοι δείχνουν στο επόμενη χαμηλότερης τάξης πλαίσιο.

3.1.1 Πρόβλεψη με μερική αντιστοίχιση σε πληροφορία Διαδικτύου

Η ακριβής πρόβλεψη μονοπατιών υπερσυνδέσμων στο Διαδίκτυο με βάση την ιστορική πληροφορία των διαδρομών πλοήγησης, μπορεί να οδηγήσει σε επίτευξη μείωσης των προσβάσεων στο παγκόσμιο ιστό, κάνοντας προανάκτηση δεδομένων Διαδικτύου. Μία σημαντική διεργασία για προανάκτηση είναι η δημιουργία ενός αποτελεσματικού μοντέλου πρόβλεψης, και μίας δομής δεδομένων, η οποία να είναι ικανή να αποθηκεύει ιστορικές πληροφορίες, οι οποίες ανακτήθηκαν από το Διαδίκτυο.

Η πρόβλεψη με μερική αντιστοίχιση είναι μία ευρέως χρησιμοποιούμενη τεχνική για προανάκτηση δεδομένων Διαδικτύου [46] [47]. Στη μέθοδο αυτή διατηρείται ένα δυναμικό μαρκοβιανό δένδρο πρόβλεψης, το οποίο διατηρεί ιστορική πληροφορία υπερσυνδέσμων και χρησιμοποιείται για εφαρμογή αποφάσεων προανάκτησης. Η μέθοδος μπορεί να χρησιμοποιηθεί για την αξιολόγηση προτύπων πλοήγησης στο Διαδίκτυο. Η συμπεριφορά περιήγησης στο Διαδίκτυο από ένα πελάτη ονομάζεται συνεδρία πελάτη και αποτελείται από μία σειρά από υπερσυνδέσμους Διαδικτύου τους οποίους επισκέπτεται ο πελάτης.

3.2 Συσχέτιση συμβάντων

Τα συμβάντα είναι μεμονωμένα ή συγκεντρωτικά μηνύματα, ή συναγερμοί, οι οποίοι περιγράφουν ή σχετίζονται με δραστηριότητες σε ένα δίκτυο [48]. Ο γενικός αυτός ορισμός καλύπτει σχεδόν όλα τα συμβάντα, τα οποία ανταλλάσσονται σε όλων των ειδών συστήματα παρακολούθησης και αξιολόγησης. Ένας ορισμός ενός συμβάντος είναι η εμφάνιση στην πηγή δεδομένων, η οποία ανιχνεύεται από ένα αισθητήρα και μπορεί να οδηγήσει στη διάδοση ενός συναγερμού. Ένας συναγερμός ορίζεται ως ένα μήνυμα από ένα στοιχείο αναλυτή σε ένα στοιχείο διαχειριστή, το οποίο περιγράφει ότι έχει γίνει

ανίχνευση ενός συμβάντος ενδιαφέροντος [48]. Ένας συναγερμός τυπικά περιέχει πληροφορία σχετικά με την ασυνήθιστη συμπεριφορά που εντοπίστηκε, καθώς επίσης και τις λεπτομέρειες της εμφάνισης.

Εκτός από τα κανονικά συμβάντα, τα οποία περιγράφουν ένα μεμονωμένο συμβάν ή συναγερμό όπως αυτό παρατηρήθηκε από το στοιχείο ανίχνευσης, η πληροφορία μεταδεδομένων διαδραματίζει ένα σημαντικό ρόλο στη συσχέτιση συμβάντων [49]. Τα μεταδεδομένα προσδιορίζουν όλη την επιπλέον πληροφορία, η οποία είναι χρήσιμη για την αξιολόγηση ενός συμβάντος. Η πληροφορία μεταδεδομένων δεν είναι μέρος της πληροφορίας μεμονωμένου συμβάντος ή άλλων συμβάντων. Ένα παράδειγμα πληροφορίας μεταδεδομένων είναι για παράδειγμα, σε ένα δίκτυο αισθητήρων, η θέση ενός αισθητήρα. Τα μεταδεδομένα συμβάλλουν σε μεγάλο βαθμό στην κατανόηση της φύσεως ενός συμβάντος. Σε πολλές εφαρμογές, αρκετά μεμονωμένα συμβάντα μπορεί να συσχετίζονται μεταξύ τους με βάση την πληροφορία μεταδεδομένων. Επίσης, τέτοιου είδους πληροφορία μπορεί να περιλαμβάνει διαχειριστική πληροφορία ή ενεργές εισόδους χρηστών. Παραδείγματα μεταδεδομένων είναι η τοπολογική πληροφορία σε ένα δίκτυο, διαχειριστική πληροφορία για συστήματα υπολογιστών και μηνύματα σφάλματος από χρήστες.

Στην επιστήμη των μαθηματικών, δύο στατιστικά τυχαίες μεταβλητές σχετίζονται, εάν υπάρχει σχέση εξάρτησης μεταξύ τους. Με την ίδια βασική έννοια, δύο συμβάντα σχετίζονται μεταξύ τους, όταν υπάρχει κάποια σύνδεση αιτιότητας. Η διαδικασία συσχέτισης συμβάντων [50] [51] έχει ως στόχο την εύρεση αυτών των συνδέσεων μεταξύ μεμονωμένων συμβάντων. Η συσχέτιση συμβάντων μπορεί να έχει ως αποτέλεσμα ένα ενιαίο συμβάν, το οποίο αποτελείται από το συνδυασμό των επιμέρους συμβάντων και ονομάζεται μετασυμβάν. Επίσης, η συσχέτιση συμβάντων μπορεί να χρησιμοποιηθεί σε εφαρμογές κατηγοριοποίησης.

Ο κύριος στόχος της διαδικασίας συσχέτισης είναι ο εντοπισμός συμβάντων, τα οποία μπορούν να χαρακτηριστούν ως περισσότερο σημαντικά, σε ένα πιθανά μεγάλο σύνολο από καταγεγραμμένα συμβάντα. Σε μεγάλο αριθμό εφαρμογών, για το χαρακτηρισμό των συμβάντων ως σημαντικά, σε κάθε συμβάν γίνεται ανάθεση μίας τιμής προτεραιότητας, η οποία εξετάζεται κατά την εκτέλεση της διαδικασίας απόφασης για αποστολή απόκρισης. Η συσχέτιση συμβάντων [50] μπορεί να οριστεί ως η διαδικασία για εδραίωση συμβάντων, έτσι ώστε να αυξάνεται η ποιότητα πληροφορίας τους, ενώ ταυτόχρονα μειώνοντας τον αριθμό των συμβάντων. Η πληροφορία μεταδεδομένων, όπως ο χρόνος, η τοποθεσία, η τοπολογία δικτύου και η διαχειριστική πληροφορία,

μπορεί να χρησιμοποιηθεί για βελτίωση της ποιότητας μεμονωμένων ή διασυνδεδεμένων συμβάντων.

Οι μέθοδοι συσχέτισης συμβάντων μπορούν να κατηγοριοποιηθούν σε δύο ομάδες, η ανάλυση με βάση τους κανόνες και η ανίχνευση ανωμαλιών. Στην ανάλυση με βάση τους κανόνες, γίνεται εφαρμογή προκαθορισμένων κανόνων ή υπογραφών σε συμβάντα, και γίνεται αλλαγή ή ταξινόμηση συμβάντων σε περίπτωση αντιστοίχισης κανόνων. Στην ανίχνευση ανωμαλιών, γίνεται καθορισμός της τυπικής συμπεριφοράς ενός συστήματος χρησιμοποιώντας μεθόδους μηχανικής μάθησης. Εάν γίνει ανίχνευση απόκλισης από την τυπική συμπεριφορά, η οποία ανιχνεύεται με υπέρβαση κάποιας τιμής κατωφλίου, γίνεται δημιουργία συμβάντων για την αναφορά της ανωμαλίας.

3.3 Μεταβλητής τάξης μαρκοβιανά μοντέλα

Τα μαρκοβιανά μοντέλα είναι ένας τυπικός τρόπος μοντελοποίησης μίας ακολουθίας από ενέργειες, οι οποίες παρακολουθούνται στο χρόνο [52]. Στην πιο απλή του μορφή, μία μαρκοβιανή αλυσίδα είναι μία στοχαστική διαδικασία, η οποία χαρακτηρίζεται με τη μαρκοβιανή ιδιότητα. Στη θεωρία πιθανοτήτων, μία στοχαστική διαδικασία είναι μία συλλογή από τυχαίες μεταβλητές, οι οποίες αντιπροσωπεύουν την εξέλιξη κάποιου συνόλου τυχαίων τιμών στο χρόνο. Μία διαδικασία, η οποία χαρακτηρίζεται ως στοχαστική έχει το χαρακτηριστικό της απροσδιοριστίας, δηλαδή ακόμα και εάν είναι γνωστή η αρχική τιμή, υπάρχει ένας μεγάλος αριθμός από πιθανές κατευθύνσεις στις οποίες η διαδικασία μπορεί να εξελιχθεί. Εάν μία στοχαστική διαδικασία έχει τη μαρκοβιανή ιδιότητα, αυτό σημαίνει ότι, με δεδομένη την κατάσταση στο παρόν, οι μελλοντικές καταστάσεις της διαδικασίας είναι ανεξάρτητες από τις παρελθοντικές καταστάσεις. Με άλλα λόγια, η περιγραφή της κατάστασης στο παρόν εκφράζει πλήρως την απαραίτητη πληροφορία, η οποία μπορεί να επηρεάσει την μελλοντική εξέλιξη της διαδικασίας. Σε κάθε χρονικό στάδιο, το σύστημα μπορεί να αλλάξει την κατάστασή του από την τρέχουσα κατάσταση σε μία άλλη κατάσταση, ή να παραμείνει στην ίδια κατάσταση. Η συμπεριφορά αυτή βασίζεται σε μία ορισμένη κατανομή πιθανοτήτων. Οι αλλαγές μεταξύ καταστάσεων μίας διαδικασίας ονομάζονται μεταβάσεις, και οι πιθανότητες οι οποίες σχετίζονται με μεταβάσεις καταστάσεων ονομάζονται πιθανότητες μετάβασης.

Οι μαρκοβιανές αλυσίδες καθορισμένης τάξης είναι μία επέκταση των στοχαστικών διαδικασιών, οι οποίες χαρακτηρίζονται με τη μαρκοβιανή ιδιότητα [53]. Στην κατηγορία

αυτή, η μελλοντική κατάσταση εξαρτάται από ένα προκαθορισμένο αριθμό παρελθοντικών καταστάσεων, ο οποίος συμβολίζεται με m. Παρόλο που η επέκταση αυτή μπορεί να είναι χρήσιμη για κάποιες εφαρμογές, υπάρχουν κάποια βασικά μειονεκτήματα στη χρήση αυτών των μοντέλων. Ένα βασικό μειονέκτημα είναι ότι μόνο μοντέλα με πολύ μικρή τάξη έχουν πρακτική αξία, αφού ο αριθμός των καταστάσεων των μαρκοβιανών αλυσίδων αυξάνεται εκθετικά όσο αυξάνεται η τάξη. Ένα ακόμα βασικό μειονέκτημα εμφανίζεται στις περιπτώσεις ακολουθιών ενεργειών, οι οποίες εκτελούνται από κάποιο χρήστη με σκοπό την επίτευξη κάποιου στόχου. Στις περιπτώσεις αυτές, η πιθανότητα της επόμενης εκτελούμενης ενέργειας δεν καθορίζεται πάντα με τον ίδιο προκαθορισμένο αριθμό από προηγούμενες ενέργειες. Συνήθως, υπάρχει ένα πλαίσιο παρελθοντικών ενεργειών μεταβλητού μήκους, το οποίο καθορίζει την κατανομή πιθανοτήτων εκτέλεσης ενεργειών χρήστη στο μέλλον.

Τα μαρκοβιανά μοντέλα μεταβλητής τάξης προέκυψαν από την ανάγκη καταγραφής μεγαλύτερων κανονικοτήτων, ενώ ταυτοχρόνως αποφεύγοντας την μεγάλη αύξηση μεγέθους, η οποία μπορεί να προκληθεί από την αύξηση της τάξης του μοντέλου. Σε αντίθεση με τα μοντέλα μαρκοβιανών αλυσίδων, όπου κάθε τυχαία μεταβλητή σε μία ακολουθία, η οποία χαρακτηρίζεται από την μαρκοβιανή ιδιότητα, εξαρτάται από ένα προκαθορισμένο αριθμό τυχαίων μεταβλητών, στα μαρκοβιανά μοντέλα μεταβλητής τάξης ο αριθμός των τυχαίων μεταβλητών μπορεί να μεταβάλλεται με βάση τη συγκεκριμένη παρατηρούμενη υλοποίηση, η οποία ονομάζεται εννοιολογικό πλαίσιο. Τα μαρκοβιανά μοντέλα μεταβλητής τάξης θεωρούν ότι σε πραγματικές συνθήκες, υπάρχουν κάποιες υλοποιήσεις καταστάσεων, οι οποίες αντιπροσωπεύονται από εννοιολογικά πλαίσια, στις οποίες κάποιες παρελθοντικές καταστάσεις είναι ανεξάρτητες των μελλοντικών καταστάσεων. Με αυτό τον τρόπο, μειώνεται αρκετά ο αριθμός των παραμέτρων του μοντέλου.

Οι αλγόριθμοι δημιουργίας μαρκοβιανών μοντέλων μεταβλητής τάξης, με βάση ένα πεπερασμένο αλφάβητο A, έχουν ως στόχο να μάθουν ένα πιθανοτικό αυτόματο πεπερασμένων καταστάσεων, το οποίο μπορεί να μοντελοποιεί ακολουθιακά δεδομένα κάποιας πολυπλοκότητας [54] [55]. Σε αντίθεση με τα M τάξης μαρκοβιανά μοντέλα, τα οποία επιχειρούν να προσεγγίσουν τις δεσμευμένες κατανομές της μορφής $P(\alpha|\beta)$, όπου $\beta \in A^N$ και $\beta \in A$, οι αλγόριθμοι μαρκοβιανών μοντέλων μεταβλητής τάξης μαθαίνουν τέτοιες δεσμευμένες κατανομές, όπου το μήκος πλαισίου $|\beta|$ μεταβάλλεται ανάλογα με τα διαθέσιμα στατιστικά στοιχεία των δεδομένων εκπαίδευσης. Συνεπώς, τα μαρκοβιανά

μοντέλα μεταβλητής τάξης βοηθούν στην καταγραφή μαρκοβιανών εξαρτήσεων μικρής και μεγάλης τάξης, με βάση τα παρατηρούμενα δεδομένα.

3.4 Μαρκοβιανές αλυσίδες

Πρόσφατες μελέτες της θεωρίας πιθανοτήτων βασίζονται σε διαδικασίες, για τις οποίες η γνώση προηγούμενων αποτελεσμάτων επηρεάζει τις προβλέψεις μελλοντικών πειραμάτων. Στις περισσότερες περιπτώσεις, όταν παρατηρείται μία ακολουθία από πειράματα, όλα τα προηγούμενα αποτελέσματα μπορούν να επηρεάσουν τις προβλέψεις για το επόμενο πείραμα. Οι διαδικασίες, στις οποίες το αποτέλεσμα ενός πειράματος μπορεί να επηρεάσει το αποτέλεσμα του επόμενου πειράματος ονομάζονται μαρκοβιανές αλυσίδες. Οι μαρκοβιανές αλυσίδες περιγράφονται ως εξής [53] [55]: λαμβάνεται υπόψη ένα σύνολο από καταστάσεις $K = \{\kappa_1, \kappa_2, \dots, \kappa_n\}$.Η διαδικασία ξεκινάει με μία από αυτές τις καταστάσεις και μετακινείται διαδοχικά από μία κατάσταση σε μία άλλη. Κάθε μετακίνηση ονομάζεται ένα βήμα. Εάν η αλυσίδα βρίσκεται στο παρόν στην κατάσταση κ_i, τότε μετακινείται στην κατάσταση κ_j στο επόμενο βήμα με μία πιθανότητα P_{ij}. Η πιθανότητα P_{ij} δεν εξαρτάται από τις καταστάσεις, στις οποίες βρισκόταν η αλυσίδα πριν την κατάσταση στο παρόν.

Οι πιθανότητες P_{ij} ονομάζονται πιθανότητες μετάβασης. Η διαδικασία μπορεί να παραμείνει στην κατάσταση στην οποία βρίσκεται στο παρόν, και αυτό μπορεί να συμβεί με πιθανότητα P_{ii}. Μία αρχική κατανομή πιθανοτήτων, η οποία ορίζεται στο K, ορίζει την κατάσταση έναρξης. Αυτό, στις περισσότερες περιπτώσεις, γίνεται καθορίζοντας μία συγκεκριμένη κατάσταση ως την κατάσταση έναρξης.

3.5 Θεωρία εξαρτήσεων

Η πιο βασική έννοια όσον αφορά τη θεωρία εξαρτήσεων είναι: μία δομή αποτελείται από δυαδικές και ασυμμετρικές σχέσεις μεταξύ στοιχείων, οι οποίες ονομάζονται εξαρτήσεις [56]. Η ιδιότητα ασυμμετρίας δημιουργεί με φυσικό τρόπο μία ιεραρχία, όπου ένα στοιχείο έχει το ρόλο της κεφαλής ή κυβερνήτη, και ένα άλλο στοιχείο έχει το ρόλο του εξαρτώμενου ή τροποποιητή. Ένα σημαντικό μέρος στη δημιουργία της θεωρίας εξάρτησης είναι η θέσπιση κριτηρίων, για επιβολή εξαρτήσεων μεταξύ των στοιχείων. Τα κριτήρια αυτά μπορούν να είναι συντακτικά, σημασιολογικά, μορφολογικά και άλλα. Στη

βιβλιογραφία [57] επισημαίνεται ότι η έννοια της εξάρτησης μπορεί εύκολα να επεκταθεί στο πεδίο της μορφολογίας, συνδέοντας στοιχεία χρησιμοποιώντας σχέσεις εξάρτησης.

3.5.1 Γράφος εξαρτήσεων

Ένας γράφος εξαρτήσεων Γ ορίζεται ως $\Gamma = (K, A)$ όπου K είναι ένα σύνολο από κόμβους ή κορυφές εξαρτήσεων και A είναι ένα σύνολο από κατευθυνόμενες ακμές εξάρτησης με ετικέτα, οι οποίες είναι υποσύνολα του K με δύο στοιχεία [58]. Κάθε ακμή μπορεί να αναπαρασταθεί ως ένα διατεταγμένο ζεύγος κορυφών εξάρτησης. Αυτά τα ζεύγη κατασκευάζουν μία λίστα, η οποία ονομάζεται λίστα συχνοτήτων. Μία ακμή εξάρτησης $\alpha = (x, y)$ όπου $\alpha \in A$ χαρακτηρίζεται ως κατευθυνόμενη από το x στο y, όπου ο κόμβος x ονομάζεται κόμβος - τροποποιητής και ο κόμβος y ονομάζεται κόμβος – κεφαλή, όταν ισχύει $x \in K, y \in K$ και $x \neq y$. Δύο ακμές ονομάζονται εφαπτόμενες, όταν αυτές διαμοιράζονται μία κοινή κορυφή ή κόμβο, και οι ακμές αυτές είναι γειτονικές σε αυτή την κορυφή ή κόμβο. Με παρόμοιο τρόπο, δύο κορυφές ή κόμβοι ονομάζονται γειτονικοί ή εφαπτόμενοι εάν διαμοιράζονται μία κοινή ακμή.

Είναι δυνατό να γίνει εφαρμογή διαφόρων λειτουργιών γράφων σε γράφους εξαρτήσεων, όπως για παράδειγμα εξαγωγή συντομότερου μονοπατιού και μείωση αριθμού κορυφών ή κόμβων. Το πρόβλημα εύρεσης συντομότερου μονοπατιού σε ένα γράφο εξαρτήσεων είναι το πρόβλημα εύρεσης ενός μονοπατιού μεταξύ δύο κορυφών ή κόμβων σε γράφο, με τέτοιο τρόπο ώστε η απόσταση μεταξύ αυτών των κορυφών ή κόμβων να ελαχιστοποιείται. Η απόσταση μεταξύ δύο κορυφών ή κόμβων ορίζεται ως το άθροισμα των βαρών των ακμών που αποτελούν μέρος του μονοπατιού. Σε κάποιες περιπτώσεις, θεωρείται ότι το βάρος σε όλες τις ακμές ισούται με ένα. Τυπικοί αλγόριθμοι για επίλυση του προβλήματος συντομότερου μονοπατιού είναι ο αλγόριθμος Μπέλμαν – Φορντ [59] [60] [61] και ο αλγόριθμος του Ντάικστρα [62]. Ο αλγόριθμος του Ντάικστρα λειτουργεί για πολλές περιπτώσεις κατευθυνόμενων γράφων με μη αρνητικές τιμές βαρών. Συνεπώς, οι γράφοι εξάρτησης είναι κατάλληλοι για ανάλυση με τον αλγόριθμο του Ντάικστρα.

Μία ακόμη σημαντική λειτουργία γράφων είναι η σύμπτυξη κορυφών ή κόμβων. Η σύμπτυξη κορυφών πραγματοποιείται σε ένα υποσύνολο των κορυφών ενός γράφου. Σε ένα γράφο εξαρτήσεων Γ, το αποτέλεσμα της σύμπτυξης δύο κορυφών x και y είναι μία νέα κορυφή z, η οποία αντικαθιστά τους κόμβους x και y με τέτοιο τρόπο, ώστε το z να

είναι γειτονικό με όλες τις κορυφές με τις οποίες οι κορυφές x και y είναι γειτονικές. Εάν οι κορυφές ή κόμβοι x και y συνδέονται με μία ακμή, τότε η ακμή αυτή αφαιρείται. Η διαδικασία αυτή ονομάζεται σύμπτυξη ακμής.

3.5.2 Δένδρο εξαρτήσεων

Ένας γράφος εξαρτήσεων Γ είναι ένα δένδρο εξαρτήσεων [63], εάν και μόνο εάν:

- Πρέπει να υπάρχει ένα ανεξάρτητο στοιχείο: δεν υπάρχει ακμή $\alpha \rightarrow \beta$ τέτοια ώστε $\beta = 0$, όπου α και β είναι κόμβοι του γράφου.
- Οι δομές εξαρτήσεων πρέπει να είναι συνδεδεμένες: για όλους τους κόμβους του γράφου, υπάρχει μία ακμή $\alpha \rightarrow* \beta$ ή $\beta \rightarrow* \alpha$.
- Κάθε εξαρτώμενο στοιχείο πρέπει να έχει τουλάχιστον μία κεφαλή: εάν ισχύει $\alpha \rightarrow \beta$, τότε δεν υπάρχει ακμή $\gamma \rightarrow \beta$ και $\gamma \neq \alpha$.

Ο πρώτος περιορισμός δηλώνει ότι πρέπει να υπάρχει ένα στοιχείο χωρίς στοιχείο κεφαλή, το οποίο είναι ένα ειδικό στοιχείο με το μηδενικό δείκτη. Το ειδικό αυτό στοιχείο είναι η ρίζα του γράφου εξαρτήσεων. Ο δεύτερος περιορισμός δηλώνει ότι υπάρχει ένα μονοπάτι μεταξύ οποιονδήποτε δύο στοιχείων σε ένα γράφο εξαρτήσεων. Ο τρίτος περιορισμός δηλώνει ότι κανένα στοιχείο δεν μπορεί να εξαρτάται από περισσότερα από ένα άλλα στοιχεία. Οι τρεις αυτοί περιορισμοί συνεπάγουν μία επιπλέον σημαντική ιδιότητα του τύπου των γράφων εξαρτήσεων που επιτρέπονται. Ο γράφος εξαρτήσεων είναι ακυκλικός, δηλαδή εάν ισχύει $\alpha \rightarrow* \beta$, τότε δεν υπάρχει ακμή $\beta \rightarrow \alpha$. Ο δεύτερος και ο τρίτος περιορισμός δεν είναι επαρκείς για να προκύψουν ακυκλικοί γράφοι, αλλά η ανεξαρτησία του στοιχείου με μηδενικό δείκτη, σύμφωνα με το πρώτο περιορισμό, σπάει όλες τις δυνατότητες δημιουργίας ενός κύκλου. Με άλλα λόγια, οι τρεις περιορισμοί συνεπάγουν ότι ο γράφος πληροί τις προϋποθέσεις ενός δένδρου με ρίζα, με όρους της θεωρίας γράφων, και επομένως μπορεί να χαρακτηριστεί και δένδρο εξαρτήσεων.

3.6 Ο αλγόριθμος μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών

Στο πλαίσιο προανάκτησης δεδομένων Διαδικτύου, ο αλγόριθμος μερικής αντιστοίχιση [1] αντιμετωπίζει το πρόβλημα της πρόβλεψης των επερχόμενων l προσβάσεων σε

διευθύνσεις Διαδικτύου ενός πελάτη, με βάση την πληροφορία από τα προηγούμενα m βήματα. Ο αλγόριθμος διατηρεί μία δομή δεδομένων, η οποία λαμβάνει υπόψη τις προηγούμενες ακολουθίες διευθύνσεων μήκους μέχρι και $m + l$. Ο αλγόριθμος καθορίζει όλες τις υποακολουθίες στην ιστορική δομή, οι οποίες ταιριάζουν με οποιοδήποτε επίθεμα των τελευταίων m προσβάσεων με στόχο την εύρεση υποψήφιων διευθύνσεων Διαδικτύου για προανάκτηση. Χρησιμοποιείται ένα κατώφλι αποκοπής $P_{κατώφλι}$ για την απόρριψη των διευθύνσεων Διαδικτύου με μικρή πιθανότητα εμφάνισης. Προανακτούνται όλα τα στοιχεία διευθύνσεων που ταιριάζουν με αυτά τα επιθέματα και έχουν πιθανότητα η οποία υπερβαίνει το κατώφλι αποκοπής $P_{κατώφλι}$.

Σε αυτή την υποενότητα γίνεται συζήτηση ενός μεταβλητής τάξης αλγορίθμου συσχέτισης συμβάντων, με την προσαρμογή της ιδέας της μερικής αντιστοίχισης στο πλαίσιο των δεδομένων ροής πολλαπλών μεταβλητών. Ο αλγόριθμος ουσιαστικά υλοποιεί την ιδέα μεταβλητής τάξης μαρκοβιανού μοντέλου [64], στην οποία γίνεται συνδυασμός μαρκοβιανών αλυσίδων τάξης $1, \dots, m$ για την πρόβλεψη ακολουθιών συμβάντων μήκους μέχρι και m. Με παρόμοιο τρόπο, όπως και στον αρχικό αλγόριθμο μερικής αντιστοίχισης, η προσέγγιση αυτή λαμβάνει υπόψη δύο παραμέτρους, η παράμετρος m και η παράμετρος l. Η παράμετρος m είναι ο αριθμός των προηγούμενων διανυσμάτων συμβάντων πολλαπλών μεταβλητών, τα οποία λαμβάνονται υπόψη. Η παράμετρος l είναι ο αριθμός των βημάτων, τα οποία ο αλγόριθμος θα προβλέψει στο μέλλον.

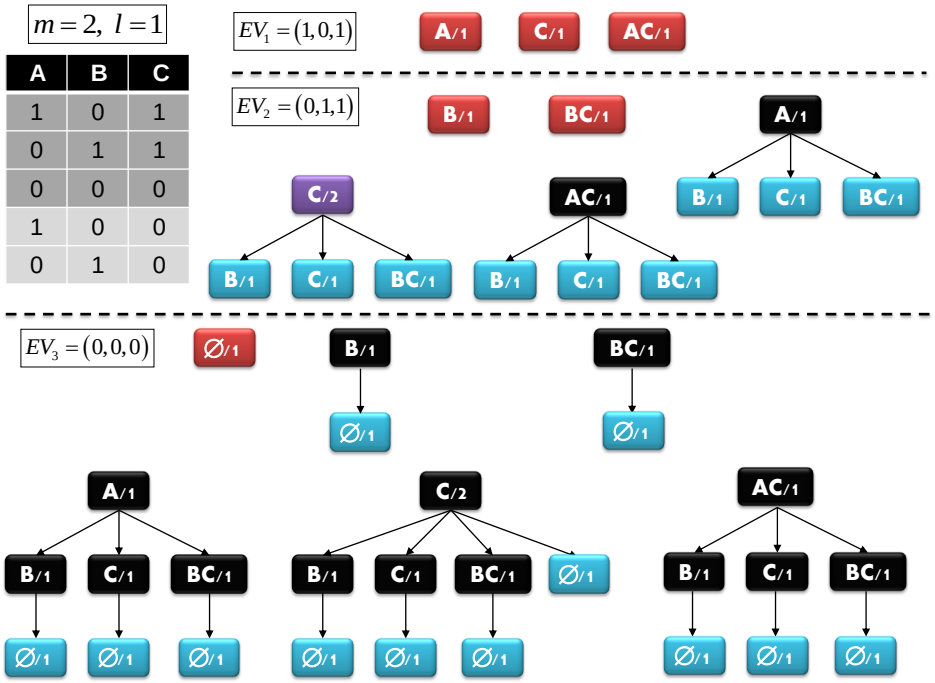
Ο αλγόριθμος λειτουργεί σε πραγματικό χρόνο, διατηρώντας μία δομή δεδομένων η οποία καταγράφει τις προηγούμενες ακολουθίες από συμβάντα μήκους μέχρι και $m + l$. Η δομή δεδομένων αυτή είναι μία συλλογή από δένδρα $\Delta = \left\{\Delta_{E_\rho}\right\}, E_\rho \in P(E)$ αντί πολλαπλών γράφων, προκειμένου να επιτευχθεί εξοικονόμηση χώρου. Η εξοικονόμηση χώρου επιτυγχάνεται λόγω του γεγονότος ότι αποθηκεύονται μόνο μία φορά κοινά προθέματα πολλαπλών ακολουθιών. Ο δείκτης $E_\rho \in E$ του κάθε δένδρου αντιπροσωπεύει το υποσύνολο συμβάντων, τα οποία σχετίζονται με το κόμβο ρίζα του δένδρου.

Σε κάθε χρονικό βήμα t, στο οποίο καταφθάνει ένα νέο διάνυσμα συμβάντων $\Delta\Sigma_t = (e_1^t, e_2^t, \dots, e_n^t) \in \{0,1\}^n$, ο αλγόριθμος ενημερώνει τη δομή δεδομένων με τρόπο τέτοιο ώστε να συμπεριλαμβάνονται όλες οι ακολουθίες συμβάτων μήκους μέχρι και $m + l$. Συνεπώς, το ύψος του κάθε δένδρου είναι κατά πολύ $h_{max} = m + l - 1$. Κάθε κόμβος δένδρου κ μπορεί να αναπαρασταθεί ως μία πλειάδα της μορφής $\kappa = \langle E_\kappa, N_\kappa^t \rangle, E_\kappa \subseteq E$,

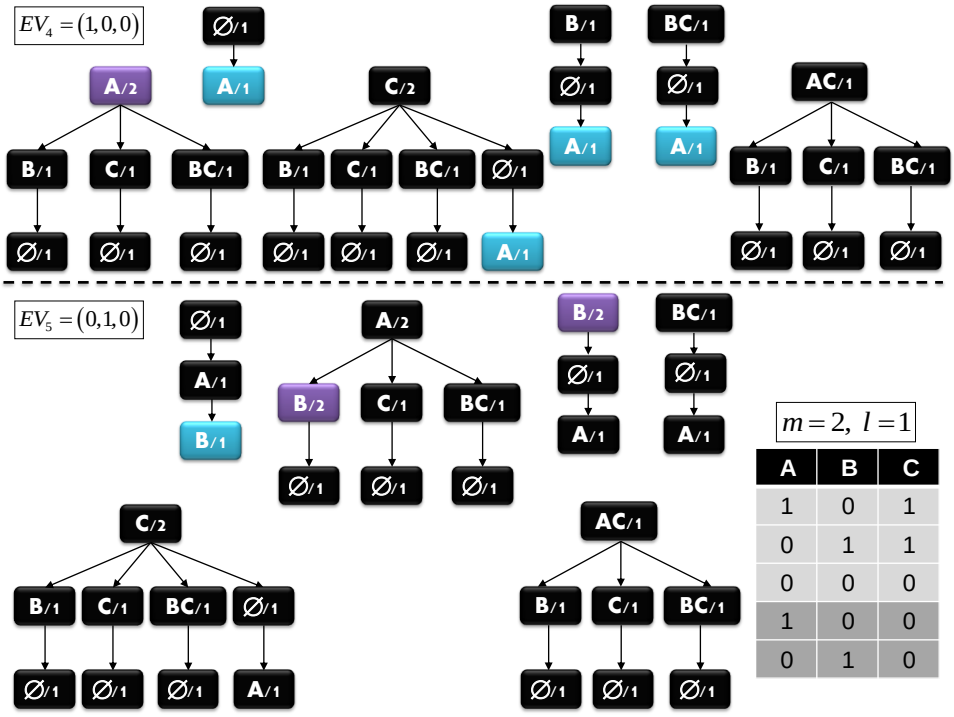
όπου E_κ είναι ένα συγκεκριμένο σύνολο από συμβάντα, τα οποία σχετίζονται με αυτό το κόμβο και N_κ^t είναι η συχνότητα του παρατηρούμενου προτύπου, ξεκινώντας από το κόμβο ρίζα μέχρι το τρέχον χρονικό βήμα t. Σε κάθε βήμα, γίνεται ενημέρωση των συχνοτήτων που αναφέρονται σε ακολουθίες των τελευταίων $m + l$ βημάτων. Οι κόμβοι αντιπροσωπεύουν συνδυασμούς συμβάντων, τα οποία εμφανίστηκαν στο ίδιο χρονικό βήμα. Συνεπώς, πολλαπλοί κόμβοι μπορεί να αντιστοιχούν στο ίδιο χρονικό βήμα. Ο λόγος για την απόφαση αυτή είναι η πρακτική ανάγκη διατήρησης διαφορετικών συχνοτήτων για τα ίδια υποπρότυπα, τα οποία εμφανίζονται μέσα σε μεγαλύτερα πρότυπα συμβάντων. Για παράδειγμα, το υποπρότυπο $B\Gamma$ μπορεί να εμφανιστεί μέσα στα πρότυπα $AB\Gamma$ και $\Gamma B\Gamma$.

Ένα ενδεικτικό παράδειγμα του αλγορίθμου με $m = 2$ και $l = 1$ παρουσιάζεται στην Εικόνα 7 για τα βήματα 1-3 και στην Εικόνα 8 για τα βήματα 4-5. Τα διανύσματα συμβάντων αναφέρονται σε τρία είδη συμβάντων A, B, και C. Σε κάθε βήμα, τα κόκκινα ορθογώνια απεικονίζουν κόμβους ρίζα για νέα δένδρα, τα οποία πρέπει να κατασκευαστούν. Τα μαύρα ορθογώνια απεικονίζουν κόμβους, οι οποίοι παραμένουν αμετάβλητοι σε σχέση με το προηγούμενο βήμα. Τα μωβ ορθογώνια απεικονίζουν κόμβους, οι οποίοι υπήρχαν στα προηγούμενα βήματα αλλά πρέπει να ενημερωθούν. Οι νέοι κόμβοι φύλλα επισημαίνονται με κυανό χρώμα. Ο ακέραιος αριθμός, ο οποίος είναι δίπλα σε κάθε κόμβο, δηλώνει τη συχνότητα του παρατηρούμενου προτύπου ξεκινώντας από τη ρίζα του δένδρου.

Η εκ των προτέρων πιθανότητα ενός συνόλου συμβάντων E_ρ ενός κόμβου ρίζα ισούται με τη συχνότητα N_ρ^t του κόμβου ρίζα, διαιρεμένη με το τρέχον χρονικό βήμα και αντιπροσωπεύεται από $P_\rho^t = \frac{N_\rho^t}{t}$. Για παράδειγμα, η εκ των προτέρων πιθανότητα του συμβάντος C στο βήμα πέντε ισούται με $P_C^5 = \frac{N_C^5}{t} = \frac{2}{5}$. Μία διαδρομή μέσα σε ένα δένδρο αντιπροσωπεύει μία ακολουθία από συμβάντα που έχουν συμβεί και συνοδεύεται από μία τιμή πιθανότητας. Η πιθανότητα ενός τέτοιου προτύπου συμβάντων της μορφής $\rho, \dots, u, v$, δηλαδή μία διαδρομή μέσα σε ένα δένδρο ξεκινώντας από το κόμβο ρίζα ρ, ισούται με τη συχνότητα του κόμβου v στο τρέχον χρονικό βήμα διαιρεμένο με τη συχνότητα του κόμβου u πριν από ένα χρονικό βήμα, και αντιπροσωπεύεται από $P_{\rho,\dots,u,v}^t = \frac{N_v^t}{N_u^{t-1}}$. Για παράδειγμα, η πιθανότητα του προτύπου $A \to B \to \emptyset$ μετά το πέμπτο χρονικό βήμα ισούται με $P_{A\to B\to\emptyset}^5 = \frac{N_\emptyset^5}{N_B^4} = 1$.



Εικόνα 8: Παράδειγμα μεταβλητής τάξης συσχέτισης συμβάντων - Βήματα 1-3



Εικόνα 9: Παράδειγμα μεταβλητής τάξης συσχέτισης συμβάντων - Βήματα 4-5

Εάν η δομή διατηρεί όλους τους πιθανούς συνδυασμούς από n διαφορετικών τύπων συμβάντα, ο αριθμός των πιθανών δένδρων στη χειρότερη περίπτωση είναι $\sum_{i=1}^{n}\binom{n}{i} = 2^n$. Εάν ληφθούν υπόψη μόνο οι συνδυασμοί, οι οποίοι αποτελούνται κατά

μέγιστο από κ συμβάντα, με κ να είναι μία σταθερή τιμή, τότε το άνω όριο του αριθμού των δένρων μειώνεται σε $\sum_{i=1}^{\kappa}\binom{n}{i} = O(n^{\kappa})$. Επιπλέον, σύμφωνα με την ίδια υπόθεση, ο μέγιστος αριθμός κόμβων που περιλαμβάνονται σε κάθε δένδρο, δηλαδή στην περίπτωση που όλα τα δένδρα είναι πλήρη και ολοκληρωμένα, είναι $\sum_{i=1}^{h_{max}}(\sum_{j=1}^{\kappa}\binom{n}{j})^{i} = \frac{1-n^{\kappa(m+l-1)}}{1-n^{\kappa}} = O(n^{m+l-1})$ όπου $h_{max} = m + l - 1$ είναι το μέγιστο ύψος του κάθε δένδρου. Πρακτικά, αυτό σημαίνει ότι ο συνολικός αριθμός από κόμβους που πρέπει να ενημερώνονται σε κάθε βήμα για όλα τα δένδρα είναι πολυωνυμικός, δεδομένου ότι ο αλγόριθμος θα εξετάσει κατά μέγιστο συνδυασμούς από κ τύπους συμβάντων. Όλες οι συχνότητες μπορούν να ενημερώνονται συσσωρευτικά με την επέκταση της ροής συμβάντων.

4. ΠΡΟΒΛΕΨΗ ΣΥΜΒΑΝΤΩΝ ΜΕ ΠΙΘΑΝΟΤΙΚΗ ΧΡΟΝΙΚΗ ΣΥΛΛΟΓΙΣΤΙΚΗ

Η ικανότητα της επίσημης έκφρασης των εξαρτήσεων μεταξύ δεδομένων συμβάντων πολλαπλών μεταβλητών και της συλλογιστικής πάνω στις διαφορετικές καταστάσεις του συστήματος με την πάροδο του χρόνου έχουν μεγάλη σημαντικότητα. Αυτό έγκειται στο γεγονός ότι, τα παραπάνω μπορούν να παρέχουν ακριβείς προβλέψεις για τη μελλοντική συμπεριφορά του συστήματος και να διευκολύνουν την εποπτεία του συστήματος από εμπειρογνώμονες. Σε συνέχεια του συστήματος συσχετισμού συμβάντων που παρουσιάστηκε προηγουμένως, η ενότητα αυτή έχει ως στόχο να παρέχει ένα κοινό πλαίσιο για την αναπαράσταση της εσωτερικής δυναμικής μίας χρονοσειράς συμβάντων. Οι τυπικές πιθανολογικές λογικές και εργαλεία [65] [66] παρέχουν τα βασικά εργαλεία για τη συλλογιστική πάνω σε αβέβαια δεδομένα, επιτρέποντας το σχολιασμό βασικών γεγονότων με μία τιμή πιθανότητας και τη χρήση κανόνων. Ωστόσο, στις περισσότερες των περιπτώσεων, αυτό δεν είναι αρκετό για την έκφραση χρονικών συσχετίσεων μεταξύ των παρατηρούμενων προτύπων. Με σκοπό την αντιπροσώπευση αβέβαιων δεδομένων και χρονικών εξαρτήσεων, στη βιβλιογραφία έχουν προταθεί πιθανοτικά χρονικά λογικά προγραμματιστικά παραδείγματα [67] [68] [69]. Τέτοιες προσεγγίσεις επεκτείνουν το συντακτικό και τη σημασιολογία των πιθανοτικών λογικών προγραμμάτων, επιτρέποντας την συλλογιστική για πιθανότητες σημείων πάνω σε χρονικά διαστήματα, με τη χρήση πιθανοτικών χρονικών κανόνων.

4.1 Πιθανοτικός Χρονικός Λογικός Προγραμματισμός

Ένας απλοποιημένος συμβολισμός για την αναπαράσταση ενός πιθανοτικού χρονικού λογικού κανόνα [67] είναι $X \rightarrow \Psi$: $[t, P]$ όπου X και Ψ είναι τύποι, οι οποίοι αποτελούνται από άτομα και τυπικές λογικές πράξεις όπως σύζευξη, διάζευξη και άρνηση, ενώ ο τύπος B περιέχει μία τιμή πιθανότητας P και μία μονάδα χρόνου t. Ο κανόνας αυτός δηλώνει ότι αν ο τύπος A που περιλαμβάνεται στο σώμα του κανόνα είναι αληθής σε καθορισμένο χρόνο, τότε ο τύπος B στην κεφαλή του κανόνα είναι επίσης αληθής με πιθανότητα P μετά από t μονάδες χρόνου ή χρονικά διαστήματα [67]. Πρακτικά, ο παραπάνω τύπος λογικών εκφράσεων επιτρέπει την έκφραση λογικών πράξεων σε άτομα που έχουν ως αποτέλεσμα την πιθανοτική αλήθεια άλλων ατόμων εντός συγκεκριμένου χρονικού διαστήματος.

Επιπλέον, τα πιθανοτικά χρονικά λογικά προγράμματα που παρουσιάζονται στο [67] επιτρέπουν τη χρήση περιορισμών ακεραιότητας που δεν πρέπει να παραβιάζονται σε όλο το χρόνο εκτέλεσης του λογικού προγράμματος. Συγκεκριμένα, οι συγγραφείς προτείνουν δύο είδη περιορισμών ακεραιότητας, περιορισμοί μεγέθους μπλοκ και περιορισμοί εμφανίσεων. Οι περιορισμοί μεγέθους μπλοκ χρησιμοποιούνται για να δηλώσουν ότι ένα άτομο X δεν μπορεί να είναι διαδοχικά αληθές περισσότερες από ένα αριθμό φορές. Οι περιορισμοί εμφανίσεων δηλώνουν ότι ένα άτομο X πρέπει να είναι αληθές για ένα αριθμό φορών, οι οποίες ανήκουν σε ένα διάστημα.

Σε αυτή την υποενότητα περιγράφεται το πρότυπο πιθανοτικού χρονικού λογικού προγραμματισμού, ως ένα εργαλείο για την αντιπροσώπευση προτύπων συμβάντων, τα οποία προέκυψαν από ένα σύστημα συσχέτισης συμβάντων, και τη διαμόρφωση των πιθανοτικών χρονικών εξαρτήσεων μεταξύ τους. Συγκεκριμένα, η εμφάνιση ενός τύπου συμβάντος X_i αντιπροσωπεύεται σε κάθε χρονικό βήμα t με το χρονικό τύπο X_i^t. Ο πιθανοτικός χρονικός τύπος, ο οποίος αποτελείται από άτομα X_i^t με πιθανότητα P αντιπροσωπεύεται από $X_i^t \ / \ P$. Οι χρονικές εξαρτήσεις μεταξύ συμβάντων αντιπροσωπεύονται μέσω των πιθανοτικών χρονικών κανόνων. Συνεπώς, ένα πιθανοτικό χρονικό λογικό πρόγραμμα μπορεί να εκφραστεί ως ένα πεπερασμένο σύνολο από πιθανοτικούς χρονικούς τύπους της μορφής $X_i^t \ / \ P$, χρονικούς πιθανοτικούς κανόνες και περιορισμούς ακεραιότητας.

Στο [67] οι συγγραφείς δείχνουν ότι η προσέγγιση τους μπορεί να χειριστεί στοχαστικές διαδικασίες, οι οποίες δεν συμμορφώνονται με τη μαρκοβιανή υπόθεση, με τέτοιο τρόπο ώστε οι μελλοντικές καταστάσεις του κόσμου να μπορούν να εξαρτώνται από περισσότερες καταστάσεις από την τρέχουσα κατάσταση. Χωρίς απώλεια της γενικότητας, μπορεί να γίνει αντιπροσώπευση μεγαλύτερης τάξης εξαρτήσεων συμβάντων με παρόμοιο τρόπο. Για παράδειγμα, το πρότυπο $A \rightarrow B \rightarrow \emptyset$ μετά το βήμα πέντε στην Εικόνα 9 μπορεί να αναπαρασταθεί ως $A_{t-1} \cap B_t \rightarrow \emptyset$: [1,1] όπου ο τύπος A_{t-1} και ο τύπος B_t χρησιμοποιούνται για την αντιπροσώπευση του τύπου A και του τύπου B σε διαδοχικά χρονικά βήματα. Ο κανόνας δηλώνει ότι εάν το A συμβαίνει ένα βήμα πριν συμβεί το B, τότε το $\emptyset$ είναι αληθές στο επόμενο χρονικό βήμα. Χρησιμοποιώντας την ίδια συντακτική, μπορεί επίσης να γίνει ρητή αναφορά σε εξαρτήσεις συμβάντων εντός χρονικών διαστημάτων.

4.2 Κανόνες συσχέτισης

Ένας κανόνας συσχέτισης είναι μία συσχέτιση της μορφής $X \Rightarrow \Psi$, όπου η παρουσία του X είναι πιθανό να συνεπάγει την παρουσία του Ψ [70] [71]. Οι κανόνες συσχέτισης μπορούν να κατηγοριοποιηθούν σε δύο τύπους, τους παραδοσιακούς κανόνες συσχέτισης και τους χρονικούς κανόνες συσχέτισης.

Μία σημαντική πτυχή των κανόνων συσχέτισης είναι ότι οι κρυμμένες συσχετίσεις δεν είναι εγγενείς στα δεδομένα, και δεν αντιπροσωπεύουν αιτιότητα μεταξύ των στοιχείων. Οι κανόνες συσχέτισης ανιχνεύουν μόνο κοινή χρήση των στοιχείων στο σύνολο δεδομένων. Ο τρόπος με τον οποίο αντιπροσωπεύονται τα δεδομένα έχει επίδραση στους κανόνες συσχέτισης που προκύπτουν. Η διαδικασία εξόρυξης κανόνων συσχέτισης έχει συνήθως την απαίτηση τα δεδομένα να είναι μοντελοποιημένα σε συναλλαγές, όπου κάθε συναλλαγή αποτελείται από ένα ή περισσότερα στοιχεία.

4.2.1 Μετρικές υποστήριξης και εμπιστοσύνης

Οι κανόνες συσχέτισης χαρακτηρίζονται από μετρικές, οι οποίες βοηθούν στη βαθμολόγηση τους, οι μετρικές της υποστήριξης (support) και της εμπιστοσύνης (confidence) [70] [72]. Η υποστήριξη ενός κανόνα συσχέτισης $X \Rightarrow \Psi$ ορίζεται ως το ποσοστό των συναλλαγών, στο οποίο τόσο το στοιχείο X όσο και το στοιχείο Ψ είναι υπαρκτά. Με άλλα λόγια, η υποστήριξη είναι η πιθανότητα της ένωσης των στοιχειοσυνόλων X και Ψ, και συμβολίζεται $P(X \cup \Psi)$. Η τιμή υποστήριξης ενός κανόνα συσχέτισης παρέχει μία ένδειξη της σπανιότητας ή ιδιαιτερότητας του κανόνα. Το όριο τιμών της υποστήριξης ενός κανόνα είναι $[0,1]$. Μία τιμή υποστήριξης, η οποία είναι κοντά στη μονάδα δείχνει ότι ο κανόνας είναι πάντα παρόν, ενώ μία τιμή υποστήριξης, η οποία είναι κοντά στο κατώτατο όριο δείχνει ότι ο κανόνας είναι μη συχνός, το οποίο μπορεί να σημαίνει ότι ο κανόνας δεν έχει ιδιαίτερο ενδιαφέρον και δυναμική ή είναι αποτέλεσμα θορύβου στο σύνολο δεδομένων.

Η εμπιστοσύνη ενός κανόνα συσχέτισης είναι ο λόγος του αριθμού των συναλλαγών, οι οποίες περιέχουν τόσο το σώμα όσο και την κεφαλή του κανόνα, προς τον αριθμό των συναλλαγών που περιέχουν το σώμα του κανόνα. Αυτό είναι ίσο με την πιθανότητα να είναι αληθής η κεφαλή του κανόνα, δεδομένου ότι το σώμα του κανόνα είναι αληθής.

$$Εμπιστοσύνη(X \Rightarrow \Psi) = \frac{P(X \cup \Psi)}{P(X)}$$

Η μετρική εμπιστοσύνης ενός κανόνα συσχέτισης δίνει μία ένδειξη της ισχύος ενός κανόνα. Μία τιμή εμπιστοσύνης η οποία είναι κοντά στο ένα, δείχνει ότι εάν το στοιχείο X είναι παρόν σε μία συναλλαγή, τότε πάντα είναι παρόν και στο στοιχείο Ψ, ενώ μία τιμή εμπιστοσύνης η οποία είναι κοντά στο κατώτατο όριο, δείχνει ότι η παρουσία του κανόνα μπορεί να είναι τυχαία, και συνεπώς ο κανόνας μπορεί να μην παρουσιάζει ενδιαφέρον.

4.2.2 Παραδοσιακοί κανόνες συσχέτισης

Οι παραδοσιακοί κανόνες συσχέτισης είναι κανόνες της μορφής $X \Rightarrow \Psi$ με υποστήριξη $\upsilon\%$ και εμπιστοσύνη $\varepsilon\%$, όπου X και Ψ είναι στοιχεία, τα οποία υπάρχουν στην ίδια συναλλαγή [70] [71]. Ένα παράδειγμα τέτοιου κανόνα μετά την εφαρμογή εξόρυξης κανόνων συσχέτισης σε δεδομένα συναλλαγών από μία υπεραγορά μπορεί να είναι $\{πατάτες\} \Rightarrow \{μπιφτέκια\}$ με υποστήριξη 40% και εμπιστοσύνη 80%. Ο κανόνας αυτός μεταφράζεται ως: το 80% όλων των ατόμων που αγοράζουν πατάτες, αγοράζουν επίσης και μπιφτέκια. Αυτού του είδους κανόνες συσχέτισης είναι χρήσιμοι για εφαρμογές όπως η διαφήμιση και η τοποθέτηση προϊόντων στις υπεραγορές.

Η διαδικασία εύρεσης κανόνων συσχέτισης μπορεί να χωριστεί σε δύο υποδιαδικασίες, η εύρεση όλων των μεγάλων στοιχειοσυνόλων και η παραγωγή κανόνων συσχέτισης από τα μεγάλα αυτά στοιχειοσύνολα. Ένα στοιχειοσύνολο ονομάζεται μεγάλο ή συχνό εάν ο αριθμός των εμφανίσεων, ή αλλιώς η μετρική της υποστήριξης, είναι πάνω από ένα προκαθορισμένο ελάχιστο κατώφλι υποστήριξης. Η υποδιαδικασία παραγωγής κανόνων συσχέτισης θεωρείται σχετικά απλή διαδικασία, σε σύγκριση με την υποδιαδικασία εύρεσης μεγάλων στοιχειοσυνόλων. Ο αλγόριθμος Apriori [70] είναι ένας ευρέως γνωστός αλγόριθμος, ο οποίος χρησιμοποιείται στη υποδιαδικασία εύρεσης μεγάλων στοιχειοσυνόλων.

4.2.3 Χρονικοί κανόνες συσχέτισης

Οι χρονικοί κανόνες συσχέτισης [73] [74] [75] διαφέρουν από τους παραδοσιακούς κανόνες συσχέτισης, στο γεγονός ότι προσπαθούν να μοντελοποιήσουν χρονικές σχέσεις

στα δεδομένα. Υπάρχουν πολλές κατηγορίες χρονικών κανόνων συσχέτισης, οι διασυναλλαγικοί κανόνες, οι κανόνες επεισοδίων, οι εξαρτήσεις τάσεων, οι ακολουθιακοί κανόνες συσχέτισης και οι ημερολογιακοί κανόνες συσχέτισης.

Οι παραδοσιακοί κανόνες συσχέτισης λαμβάνουν υπόψη στοιχεία τα οποία εμφανίζονται μαζί στις ίδιες συναλλαγές, και επομένως μπορούν να χαρακτηριστούν ως ενδοσυναλλαγικοί κανόνες συσχέτισης. Υπάρχουν ωστόσο περιπτώσεις, στις οποίες θα ήταν χρήσιμο να υπάρχουν κανόνες συσχέτισης οι οποίοι να εκτείνονται σε ένα εύρος συναλλαγών. Οι αλγόριθμοι για εξόρυξη διασυναλλαγικών κανόνων συσχέτισης συνήθως χρησιμοποιούν ένα χρονικό παράθυρο. Το χρονικό αυτό παράθυρο καθορίζει το μέγιστο αριθμό συναλλαγών, στις οποίες ο κανόνας συσχέτισης μπορεί να εκτείνεται. Ένα χρονικό παράθυρο με μηδενική τιμή σημαίνει ότι ο αλγόριθμος βρίσκει μόνο κανόνες συσχέτισης οι οποίοι περιέχουν στοιχειοσύνολα, τα οποία εμφανίζονται στην ίδια συναλλαγή. Συνεπώς, οι ενδοσυναλλαγικοί κανόνες μπορούν να θεωρηθούν ως μία ειδική περίπτωση των διασυναλλαγικών κανόνων. Ένα παράδειγμα ενός διασυναλλαγικού κανόνα συσχέτισης είναι $\{πατάτες, 0\} \Rightarrow \{μπιφτέκια, 1\}$ με υποστήριξη 20% και εμπιστοσύνη 90%, το οποίο σημαίνει ότι το 90% των ατόμων που αγοράζουν πατάτες, αγοράζουν μπιφτέκια την επόμενη μέρα, ή κάποια άλλη χρονική στιγμή η οποία καθορίζεται από τη χρονική μονάδα.

4.2.4 Κανόνες επεισοδίων

Στη βιβλιογραφία [73] [76] έχει γίνει ανάπτυξη εξειδικευμένων διαδικασιών για ακολουθίες συμβάντων, οι οποίες προσφέρουν ένα ευρύτερο πλαίσιο και δυναμική έκφρασης στους προκύπτοντες κανόνες. Ένα συχνά χρησιμοποιούμενο πλαίσιο είναι οι κανόνες επεισοδίων [73] [77], οι οποίοι είναι επέκταση των κανόνων συσχέτισης, και συγκεκριμένα των χρονικών κανόνων συσχέτισης. Η προσέγγιση αυτή ενσωματώνει τους μηχανισμούς, οι οποίοι είναι απαραίτητοι για την επέκταση της εξόρυξης συναλλαγικών κανόνων συσχέτισης με τη διάσταση του χρόνου μέσω της έννοιας των επεισοδίων.

Στο πλαίσιο αυτό, η είσοδος είναι μία ακολουθία από συμβάντα, όπου κάθε συμβάν έχει συσχετιζόμενη χρονική πληροφορία εμφάνισης. Τυπικά, τα δεδομένα είναι ένα σύνολο από συμβάντα Υ, όπου κάθε συμβάν χαρακτηρίζεται από ένα ζεύγος (T, t) όπου $T \in \Upsilon$ είναι ο τύπος του συμβάντος και t είναι ένας ακέραιος αριθμός, ο οποίος αντιπροσωπεύει το χρόνο εμφάνισης του συμβάντος. Μία ακολουθία συμβάντων α στο

Υ είναι μία τριάδα της μορφής $(\alpha, t_\varepsilon, t_\tau)$ όπου t_ε είναι ο χρόνος έναρξης και t_τ είναι ο χρόνος τερματισμού της ακολουθίας, όπου $t_\varepsilon \leq t_\tau$ είναι ακέραιοι αριθμοί, $\alpha = \langle (T_1, t_1), (T_2, t_2), \dots, (T_n, t_n) \rangle$ και $T_i \in \Upsilon$ και $t_\varepsilon \leq t_i \leq t_\tau$ για όλα τα $i = 1, \dots, n$.

Ένα βασικό πρόβλημα στην ανάλυση αυτών των ακολουθιών είναι η εύρεση συχνών επεισοδίων. Ένα επεισόδιο ορίζεται ως μία συλλογή από συμβάντα, τα οποία συμβαίνουν συχνά μαζί. Ένας γενικότερος ορισμός είναι: τα επεισόδια είναι μερικώς διατεταγμένα σύνολα συμβάντων, τα οποία μπορεί να είναι σειριακά ή παράλληλα. Ένα επεισόδιο ορίζεται ως ένα ζεύγος $(\Phi, \leq)$ όπου Φ είναι μία συλλογή από τύπους συμβάντων και $\leq$ είναι μία μερική διάταξη του Φ. Ένα επεισόδιο είναι παράλληλο, εάν η σχέση μερικής διάταξης είναι αμελητέα, ενώ ένα επεισόδιο είναι σειριακό εάν η σχέση διάταξης είναι απόλυτη. Οι κανόνες επεισοδίων δίνουν το απαιτούμενο πλαίσιο, το οποίο είναι απαραίτητο για διαχείριση της διαδικασίας εξόρυξης κανόνων σε ακολουθίες συμβάντων. Οι κανόνες επεισοδίων επίσης εισάγουν την έννοια του χρόνου στους κανόνες.

Ένας κανόνας συσχέτισης $X \to \Psi$ αποτελείται από δύο μέρη, το σώμα του κανόνα, στην περίπτωση αυτή το X, και η κεφαλή του κανόνα, στην περίπτωση αυτή το Ψ. Αυτό δίνει τέσσερεις πιθανούς συνδυασμούς του αριθμού των συμβάντων στην κεφαλή και στο σώμα του κανόνα [73] [77]. Κάθε μέρος του κανόνα μπορεί να αποτελείται από ένα μεμονωμένο συμβάν ή πολλαπλά συμβάντα. Οι πιθανές μορφές της δομής των κανόνων παρουσιάζονται στον Πίνακα 1. Υπάρχουν τέσσερις συνδυασμοί του αριθμού, μεμονωμένο ή πολλαπλά, των συμβάντων στο σώμα και στην κεφαλή του κανόνα.

Πίνακας 1: Πιθανές δομές κανόνων συσχέτισης

Δομή κανόνα	**Παράδειγμα**
$\mu\varepsilon\mu o\nu\omega\mu\acute{\varepsilon}\nu o \to \mu\varepsilon\mu o\nu\omega\mu\acute{\varepsilon}\nu o$	$X \to \Psi$
$\mu\varepsilon\mu o\nu\omega\mu\acute{\varepsilon}\nu o \to \pi o\lambda\lambda\alpha\pi\lambda\acute{\alpha}$	$X \to \Psi, Z$
$\pi o\lambda\lambda\alpha\pi\lambda\acute{\alpha} \to \mu\varepsilon\mu o\nu\omega\mu\acute{\varepsilon}\nu o$	$X, \Psi \to Z$
$\pi o\lambda\lambda\alpha\pi\lambda\acute{\alpha} \to \pi o\lambda\lambda\alpha\pi\lambda\acute{\alpha}$	$X, \Psi \to Z, \Upsilon$

Με άλλα λόγια, στην τυπική σύνταξη κανόνων συσχέτισης και κανόνων επεισοδίων, επιτρέπεται οποιοσδήποτε αριθμός από στοιχεία τόσο στην κεφαλή όσο και στο σώμα του κανόνα.

5. ΠΡΟΣΑΡΜΟΣΤΙΚΟ ΦΙΛΤΡΑΡΙΣΜΑ ΕΞΑΡΤΗΣΕΩΝ ΣΥΜΒΑΝΤΩΝ

Στην προηγούμενη ενότητα έγινε μία συζήτηση ενός διαφανούς τρόπου για την αναπαράσταση συσχέτισης συμβάντων, η οποία είναι ανεξάρτητη από το υποκείμενο σύστημα συσχέτισης. Ένα πρόσθετο βήμα στη διαχείριση συμβάντων σε πραγματικό χρόνο ή σε σχεδόν πραγματικό χρόνο λαμβάνει υπόψη το φιλτράρισμα των προκύπτοντων εξαρτήσεων. Συγκεκριμένα, η συσχέτιση συμβάντων παρέχει χρήσιμη πληροφορία για τη συμπεριφορά του συστήματος σε συνάρτηση με το χρόνο. Ωστόσο, τα περισσότερα δίκτυα αισθητήρων είναι σε μεγάλο βαθμό δυναμικά. Η δυναμική αυτή συμπεριφορά έγκειται στο ότι τα δίκτυα αισθητήρων έχουν ως στόχο την καταγραφή μεταβολών μέσα σε μη στατικά περιβάλλοντα και χώρους. Ως εκ τούτου, οι εξαρτήσεις που προκύπτουν σε κάθε βήμα μπορούν να αλλάξουν με την πάροδο του χρόνου. Στο πλαίσιο αυτό, ένας περιορισμός που χαρακτηρίζει τις τυπικές μεθόδους συσχέτισης συμβάντων είναι η μη ικανότητα τους για αναγνώριση χρονικών εξαρτήσεων. Για παράδειγμα, κανόνες οι οποίοι προέκυψαν από ένα σύστημα συσχέτισης το οποίο βασίζεται σε μαρκοβιανά μοντέλα μπορεί να αντικατοπτρίζουν εξαρτήσεις μεταξύ συμβάντων, τα οποία έλαβαν χώρα πριν από αρκετό χρονικό διάστημα και δεν είναι σύμφωνες με την πρόσφατη συμπεριφορά των συμβάντων. Μία αντίθετη περίπτωση είναι η περίπτωση ενός χωρίς μνήμη συστήματος, όπως για παράδειγμα ένα σύστημα το οποίο βασίζεται σε κυλιόμενο παράθυρο, το οποίο λαμβάνει υπόψη μόνο την πιο πρόσφατη συμπεριφορά, αγνοώντας πλήρως τις προηγούμενες εμφανίσεις συμβάντων.

Μία απλή, αλλά αποτελεσματική προσέγγιση για την επίλυση του προβλήματος καθορισμού του πότε ένας προκύπτων κανόνας ή εξάρτηση γίνεται παλιός, είναι η εφαρμογή μίας χρονικά εξαρτώμενης συνάρτησης απόσβεσης. Σε γενικές γραμμές, οι συναρτήσεις απόσβεσης αποτελούν μηχανισμούς λήθης, με βάση το χρόνο. Η βασική ιδέα είναι ότι, μία μετατόπιση στη συμπεριφορά σε μία ροή δεδομένων θα πρέπει να είναι μία σταδιακή διαδικασία. Επίσης, οι κανόνες οι οποίοι προέκυψαν πρόσφατα πρέπει να είναι μεγαλύτερης σημαντικότητας από τους κανόνες, οι οποίοι είναι πιο παλιοί. Στο ίδιο πλαίσιο, η σημαντικότητα ενός κανόνα πρέπει να μειώνεται με την πάροδο του χρόνου. Στη συνέχεια της ενότητας αυτή περιγράφεται το θεωρητικό υπόβαθρο πάνω στο οποίο βασίζεται το προσαρμοστικό φιλτράρισμα εξαρτήσεων συμβάντων και παρουσιάζονται οι συναρτήσεις που χρησιμοποιήθηκαν στα πλαίσια της εργασίας, μία γραμμική συνάρτηση απόσβεσης και μία εκθετική συνάρτηση απόσβεσης.

5.1 Χρονικά χαρακτηριστικά μοντέλου επεξεργασίας δεδομένων

Από τον ορισμό που δόθηκε προηγουμένως, οι ροές δεδομένων αποτελούνται από διατεταγμένες ακολουθίες στοιχείων. Κάθε σύνολο δεδομένων συνήθως ονομάζεται συναλλαγή. Επειδή τα δεδομένα φθάνουν με συνεχή και απεριόριστο τρόπο, οι συναλλαγές που προκύπτουν από τα δεδομένα μεταβάλλονται με το χρόνο. Υπάρχουν τρία μοντέλα επεξεργασίας δεδομένων αισθητήρων στη βιβλιογραφία [78], το μοντέλο οροσήμου, το μοντέλο απόσβεσης και το μοντέλο κυλιόμενου παραθύρου.

Στο μοντέλο οροσήμου, η ανακάλυψη γνώσης γίνεται σε ολόκληρο το ιστορικό των δεδομένων αισθητήρων από μία χρονική στιγμή, η οποία ονομάζεται ορόσημο, μέχρι την τρέχουσα χρονική στιγμή. Το μοντέλο οροσήμου δεν είναι κατάλληλο για εφαρμογές, στις οποίες υπάρχει ενδιαφέρον μόνο για την πιο πρόσφατη πληροφορία δεδομένων αισθητήρων.

Το μοντέλο απόσβεσης επεξεργάζεται δεδομένα αισθητήρων, στα οποία κάθε συναλλαγή έχει κάποια τιμή βάρους, η οποία μειώνεται με την πάροδο του χρόνου. Με αυτό τον τρόπο, οι παλαιότερες συναλλαγές συμβάλλουν λιγότερο στη διαδικασία ανακάλυψης γνώσης από ότι οι πιο πρόσφατες. Συνεπώς, το μοντέλο αυτό λαμβάνει υπόψη διαφορετικά βάρη για παλαιότερες και πιο πρόσφατες συναλλαγές. Το μοντέλο απόσβεσης είναι κατάλληλο για εφαρμογές, στις οποίες τα παλαιότερα δεδομένα έχουν κάποια επίδραση στα αποτελέσματα ανακάλυψης γνώσης, αλλά η επίδραση αυτή μειώνεται με την πάροδο του χρόνου. Η προσέγγιση προσαρμοστικού φιλτραρίσματος εξαρτήσεων συμβάντων που διερευνήθηκε στα πλαίσια της εργασίας βασίζεται στο μοντέλο απόσβεσης.

Το μοντέλο κυλιόμενου παραθύρου κάνει ανακάλυψη γνώσης με βάση κυλιόμενα παράθυρα. Γίνεται αποθήκευση και επεξεργασία μέρους των δεδομένων αισθητήρων, τα οποία βρίσκονται μέσα στα όρια ενός χρονικώς κυλιόμενου παραθύρου σε κάθε χρονική στιγμή. Το μέγεθος του κυλιόμενου παραθύρου καθορίζεται σύμφωνα με τα χαρακτηριστικά της κάθε εφαρμογής και τους πόρους του συστήματος. Τα αποτελέσματα ανακάλυψης γνώσης εξαρτάται σε απόλυτο βαθμό από πιο πρόσφατες συναλλαγές, οι οποίες εμπίπτουν χρονικά μέσα στα όρια του κυλιόμενου παραθύρου. Σε κάθε χρονική στιγμή, αποθηκεύονται οι συναλλαγές που ανήκουν μέσα στους χρονικούς περιορισμούς του κυλιόμενου παραθύρου. Με αυτό τον τρόπο, όταν μία συναλλαγή βρίσκεται εκτός του εύρους του κυλιόμενου παραθύρου, οι επιπτώσεις της στα αποτελέσματα ανακάλυψης γνώσης είναι μηδενικές.

Η επιλογή εφαρμογής μοντέλου επεξεργασίας δεδομένων εξαρτάται σε μεγάλο βαθμό από τα χαρακτηριστικά και τις ανάγκες της κάθε εφαρμογής. Μία μέθοδος που χρησιμοποιεί το μοντέλο οροσήμου μπορεί να μετατραπεί σε μέθοδο που χρησιμοποιεί το μοντέλο απόσβεσης, με την προσθήκη μίας συνάρτησης απόσβεσης στις επερχόμενες ροές δεδομένων. Επίσης, μία μέθοδος που χρησιμοποιεί το μοντέλο οροσήμου μπορεί να μετατραπεί μέθοδο που χρησιμοποιεί το μοντέλο κυλιόμενου παραθύρου, με επεξεργασία των συναλλαγών οι οποίες εμπίπτουν στους χρονικούς περιορισμούς ενός κυλιόμενου παραθύρου.

5.2 Προσαρμοστικό φιλτράρισμα για προσέγγιση μεγεθών

Υπάρχουν περιπτώσεις στις οποίες μία ακολουθία από τιμές $x_1, \dots, x_N$ οι οποίες προέρχονται από μία διαδικασία, η οποία μεταβάλλεται στο χρόνο και λαμβάνεται υπόψη η υπόθεση ότι η φύση της μεταβολής στο χρόνο είναι άγνωστη. Επίσης, λαμβάνεται υπόψη η υπόθεση ότι το ζητούμενο είναι μία καλή εκτίμηση του μέσου όρου της διαδικασίας σε κάθε χρονική στιγμή, καθώς γίνεται άφιξη των δεδομένων. Μία απλή διαδικασία υπολογισμού του μέσου όρου πάνω στην ακολουθία δεν προσφέρει σε κάθε περίπτωση μία καλή εκτίμηση της τρέχουσας τιμής του μέσου όρου. Αυτό οφείλεται στο γεγονός ότι τα δεδομένα που είναι πιο παλιά μπορεί να μην είναι τόσο χρήσιμα όσο τα πιο πρόσφατα δεδομένα. Στόχος της διαδικασίας προσαρμοστικού φιλτραρίσματος [79] [80] είναι η παροχή μίας μεθοδολογίας, η οποία να είναι ικανή στη διαχείριση της μεταβολής των δεδομένων στο χρόνο, δίνοντας περισσότερη πληροφοριακή χρησιμότητα σε δεδομένα, τα οποία είναι πιο πρόσφατα. Το προσαρμοστικό φιλτράρισμα αποτελείται από μεθόδους οι οποίες βασίζονται σε εκθετικούς παράγοντες λήθης, παράγοντες οι οποίοι εξισορροπούν παλιά δεδομένα με πιο πρόσφατα δεδομένα [81]. Εάν το ζητούμενο είναι απλές στατιστικές, όπως ο μέσος όρος, η μεθοδολογία προσαρμοστικού φιλτραρίσματος παρέχει μία αποτελεσματική ακολουθιακή εκτίμηση.

Στη συνέχεια γίνεται περιγραφή μεθόδων προσαρμοστικού φιλτραρίσματος για τις στατιστικές μέσου όρου και τυπικής απόκλισης [80] [81] χρησιμοποιώντας μία παράμετρο λ, η οποία ανήκει στο διάστημα $[0,1]$ και ονομάζεται σταθερός παράγοντας λήθης. Το πρόβλημα επιλογής της παραμέτρου είναι παρόμοιο με το πρόβλημα επιλογής παραμέτρων στους αλγορίθμους ανίχνευσης απότομων μεταβολών. Στη συνέχεια θα γίνει μία περιγραφή κατασκευής ενός ανιχνευτή απότομων μεταβολών, ο οποίος χρησιμοποιεί προσαρμοστικό φιλτράρισμα με ένα σταθερό παράγοντα λήθης.

Λαμβάνεται υπόψη η υπόθεση ότι υπάρχει μία ακολουθία από N παρατηρήσεις $x_1, \dots, x_N$. Ο μέσος όρος δείγματος $\overline{x_N}$ ορίζεται ως:

$$\overline{x_N} = \frac{1}{N} \times \sum_{i=1}^{N} x_i$$

Ο μέσος όρος δείγματος με παράγοντα λήθης $\overline{x_{N,\lambda}}$ ορίζεται ως:

$$\overline{x_{N,\lambda}} = \frac{1}{\beta_{N,\lambda}} \times \sum_{i=1}^{N} \lambda^{N-i} \times x_i$$

με συντελεστή λήθης $\lambda \in [0,1]$ όπου το βάρος $\beta_{N,\lambda}$ ορίζεται ως:

$$\beta_{N,\lambda} = \sum_{i=1}^{N} \lambda^{N-i}$$

Σε ένα μεγάλο αριθμό εφαρμογών, όταν η διαδικασία υφίσταται μία απότομη μεταβολή, ένας απλός υπολογισμός μέσου όρου δεν παρέχει μία ενημερωμένη εκτίμηση του μέσου όρου της διαδικασίας, σε σύγκριση με τις εκτιμήσεις που προήλθαν από το προσαρμοστικό φιλτράρισμα.

Ο μέσος όρος δείγματος με παράγοντα λήθης, όπως ορίστηκε προηγουμένως, μειώνει εκθετικά την τιμή βάρους σε παρατηρήσεις οι οποίες είναι πιο παλιές $(x_1, x_2, \dots)$, και επομένως δίνει μεγαλύτερη τιμή βάρους σε πιο πρόσφατες παρατηρήσεις $(\dots, x_{N-1}, x_N)$. Συνεπώς, ο μέσος όρος δείγματος με παράγοντα λήθης $\overline{x_{N,\lambda}}$ είναι μία τυχαία μεταβλητή, η οποία προσεγγίζει την τιμή μέσου όρου των παρατηρήσεων $x_1, \dots, x_N$ με τέτοιο τρόπο, ώστε να υπάρχει περισσότερο βάρος σε πιο πρόσφατες παρατηρήσεις.

Όσο η τιμή του παράγοντα λήθης λ πλησιάζει τη μηδενική τιμή, τόσο μειώνονται εκθετικά τα βάρη των πιο παλιών παρατηρήσεων, και τόσο περισσότερο βάρος αποκτούν οι πιο πρόσφατες παρατηρήσεις. Το όφελος από τη συμπεριφορά αυτή είναι ότι ο μέσος όρος με παράγοντα λήθης είναι κοντά στο μέσο όρο ενός αριθμού πιο πρόσφατων παρατηρήσεων. Ωστόσο, ένα μειονέκτημα της συμπεριφοράς αυτής είναι ότι ο μέσος όρος με παράγοντα λήθης είναι πιο ευαίσθητος σε αποκλίνουσες τιμές. Όσο η τιμή του παράγοντα λήθης λ πλησιάζει την τιμή ένα, τόσο μικρότερη τιμή βάρους υπάρχει στις πιο πρόσφατες παρατηρήσεις, και τόσο πιο κοντά είναι οι τιμές του μέσου όρου με παράγοντα λήθης $\overline{x_{N,\lambda}}$ και του μη σταθμισμένου μέσου όρου $\overline{x_N}$. Το όφελος από τη συμπεριφορά αυτή είναι ότι ο μέσος όρος με παράγοντα λήθης $\overline{x_{N,\lambda}}$ είναι λιγότερο ευαίσθητος σε αποκλίνουσες τιμές.

Μία αποκλίνουσα τιμή είναι μία παρατήρηση, η οποία βρίσκεται αρκετά εκτός του αναμενόμενου εύρους τιμών σε μία μελέτη ή πείραμα, η οποία συχνά απορρίπτεται από το σύνολο δεδομένων. Η διάκριση μεταξύ ενός σημείου απόκλισης και ενός σημείου πραγματικής απότομης μεταβολής είναι λεπτή. Ένα σημείο απότομης μεταβολής έχει τυπικά επίμονη συμπεριφορά, ενώ ένα σημείο απόκλισης έχει εφήμερη συμπεριφορά. Ένα ανοικτό θέμα μελέτης είναι η αποτελεσματική εύρεση σημείων απόκλισης και διάκριση τους από τα σημεία απότομων μεταβολών.

Στον ορισμό του μέσου όρου με παράγοντα λήθης, υπάρχουν διάφορες πιθανές επιλογές για το βάρος $\beta_{N,\lambda}$. Μία τυπική προσέγγιση είναι η επιλογή βάρους $\beta_{N,\lambda}$ έτσι ώστε ο μέσος όρος με παράγοντα λήθης $\overline{x_{N,\lambda}}$ να είναι όσο γίνεται αμερόληπτος. Με μαθηματικούς ορισμούς, εάν ληφθεί υπόψη η υπόθεση ότι για $x_1, \dots, x_N$ ισχύει $E[\overline{x_i}] = \mu$, τότε ισχύει $E\left[\overline{x_{N,\lambda}}\right] = \mu$. Εάν επιπλέον ληφθεί υπόψη η υπόθεση ότι για $i = 1, 2, \dots, N$ έχουμε $Var[x_i] = \sigma^2$, τότε ισχύει:

$$Var\left[\overline{x_{N,\lambda}}\right] = b_{N,\lambda} \times \sigma^2$$

όπου $b_{N,\lambda} = \frac{1}{(\beta_{N,\lambda})^2} \times \sum_{i=1}^{N} (\lambda^2)^{N-i} = \frac{(1-\lambda)\times(1+\lambda^N)}{(1-\lambda^N)\times(1+\lambda)}$.

Ο ορισμός του μέσου όρου δείγματος με παράγοντα λήθης προσφέρει μία καλή εκτίμηση του μέσου όρου δείγματος, λαμβάνοντας υπόψη περισσότερο βάρος στις πιο πρόσφατες παρατηρήσεις. Μπορεί να είναι χρήσιμο να υπάρχει μία καλή εκτίμηση της τυπικής απόκλισης δείγματος, λαμβάνοντας υπόψη περισσότερο βάρος στις πιο πρόσφατες παρατηρήσεις. Με παρόμοιο τρόπο, ορίζεται η μετρική τυπικής απόκλισης δείγματος με παράγοντα λήθης $\sigma_{N,\lambda}^2$. Η τυπική απόκλιση δείγματος σε παρατηρήσεις $x_1, \dots, x_N$ ορίζεται ως:

$$\sigma_N^2 = \frac{1}{N-1} \times \sum_{i=1}^{N} (x_i - \overline{x_N})^2$$

Η τυπική απόκλιση με παράγοντα λήθης $\sigma_{N,\lambda}^2$ ορίζεται ως:

$$\sigma_{N,\lambda}^2 = \frac{1}{b_{N,\lambda}} \times \sum_{i=1}^{N} \lambda^{N-i} \times (x_i - \overline{x_{N,\lambda}})^2$$

όπου $\overline{x_{N,\lambda}}$ είναι ο μέσος όρος με παράγοντα λήθης και $b_{N,\lambda}$ είναι μία τιμή βάρους, η οποία ορίζεται ως:

$$b_{N,\lambda} = \frac{2 \times \lambda \times (1 - \lambda^{N-1})}{(1-\lambda) \times (1+\lambda)}$$

5.3 Μαθηματικό υπόβαθρο

Σε αυτή την υποενότητα παρουσιάζονται οι μαθηματικοί ορισμοί, πάνω στους οποίους βασίζεται η προσέγγιση προσαρμοστικού φιλτραρίσματος εξαρτήσεων συμβάντων της εργασίας.

5.3.1 Γραμμικές και εκθετικές συναρτήσεις

Στα μαθηματικά, μία συνάρτηση χρησιμοποιείται για να αντιπροσωπεύσει την εξάρτηση μίας ποσότητας πάνω σε μία άλλη. Ένας πιο τυπικός ορισμός είναι ο ακόλουθος: μία συνάρτηση είναι ένας κανόνας, ο οποίος παίρνει κάποιους αριθμούς ως είσοδο και αναθέτει σε κάθε έναν από αυτούς έναν καθορισμένο αριθμό εξόδου. Το σύνολο όλων των αριθμών εισόδου ονομάζεται πεδίο ορισμού της συνάρτησης, και το σύνολο των αριθμών που προκύπτει στην έξοδο ονομάζεται πεδίο τιμών της συνάρτησης. Η είσοδος ονομάζεται ανεξάρτητη μεταβλητή, ενώ η έξοδος ονομάζεται εξαρτημένη μεταβλητή. Μία συνάρτηση, η οποία έχει ως είσοδο χρονικές τιμές συμβολίζεται ως f και γράφεται ως $H = f(t)$. Μία συνάρτηση μπορεί να έχει ίδια έξοδο με διαφορετικές τιμές εισόδου. Κάποια μεγέθη, όπως η ημερομηνία, ονομάζονται διακριτά, και αυτό σημαίνει ότι παίρνουν ορισμένες απομονωμένες τιμές, ενώ κάποια μεγέθη όπως ο χρόνος, ονομάζονται συνεχή και μπορούν να πάρουν οποιαδήποτε τιμή.

Εάν το πεδίο ορισμού μίας συνάρτησης δεν έχει καθοριστεί, συνήθως λαμβάνεται υπόψη να είναι το μεγαλύτερο δυνατό σύνολο των πραγματικών αριθμών. Για παράδειγμα, λαμβάνεται υπόψη ότι το πεδίο ορισμού της συνάρτησης $f(x) = 10 \times x$ είναι όλοι οι πραγματικοί αριθμοί. Ωστόσο, το πεδίο ορισμού της συνάρτησης $f(x) = \frac{10}{x}$ είναι όλοι οι πραγματικοί αριθμοί, εκτός από τη μηδενική τιμή. Μερικές φορές γίνεται περιορισμός του πεδίου ορισμού, έτσι ώστε να είναι μικρότερο από το μεγαλύτερο δυνατό σύνολο των πραγματικών αριθμών. Για παράδειγμα, εάν μία συνάρτηση υπολογίζει το εμβαδόν μίας γεωμετρίας και λαμβάνει ως είσοδο το μήκος πλευρών της, η συνάρτηση αυτή έχει ως πεδίο ορισμού μη αρνητικές τιμές.

5.3.1.1 Γραμμική συνάρτηση

Μία συνάρτηση είναι γραμμική εάν η κλίση της, ή αλλιώς ο ρυθμός αλλαγής της, είναι ίδιος σε κάθε σημείο. Ο ρυθμός αλλαγής μίας συνάρτησης, η οποία δεν είναι γραμμική, ποικίλει από σημείο σε σημείο. Εάν μία συνάρτηση $y = f(t)$ αυξάνεται με την πάροδο του χρόνου, τότε η συνάρτηση f είναι μία αυξανόμενη συνάρτηση. Εάν η συνάρτηση $y = f(t)$ μειώνεται με την πάροδο του χρόνου, τότε η συνάρτηση f είναι μία φθίνουσα συνάρτηση. Ένα παράδειγμα αυξανόμενης συνάρτησης, η οποία είναι γραμμική είναι $y = \alpha \times t + \beta$, και ένα παράδειγμα φθίνουσας συνάρτησης είναι $y = \alpha \times t - \beta$.

Μία γραμμική συνάρτηση έχει τη μορφή $y = f(x) = \alpha \times x + \beta$. Η γραφική παράσταση μίας γραμμικής συνάρτησης είναι μία ευθεία, σύμφωνα με την οποία το α είναι η κλίση ή ρυθμός αλλαγής της συνάρτησης y ως προς x, ενώ το β είναι το κατακόρυφο σημείο τομής, δηλαδή η τιμή της συνάρτησης y όταν το x έχει μηδενική τιμή. Εάν η κλίση α έχει μηδενική τιμή, ο γράφος της συνάρτησης είναι μία οριζόντια γραμμή $y = \beta$. Εάν ένας πίνακας περιέχει τιμές x και y, μπορεί να διαπιστωθεί ότι οι τιμές του πίνακα προέρχονται από γραμμική συνάρτηση $y = \alpha \times x + \beta$, εάν η διαφορά στις τιμές y είναι ίδια για τιμές x ίσης διαφοράς.

Εξισώσεις της μορφής $y = \alpha \times x + \beta$, όπου οι σταθερές α και β μπορούν να πάρουν διάφορες τιμές, δημιουργούν μία οικογένεια από συναρτήσεις. Όλες οι εξισώσεις μίας οικογένειας συναρτήσεων μοιράζονται κάποιες ιδιότητες. Στην περίπτωση αυτή, όλες οι εξισώσεις αναπαριστώνται σε γραφική παράσταση ως μία ευθεία γραμμή. Οι σταθερές α και β ονομάζονται παράμετροι. Όσο μεγαλώνει το μέγεθος της παραμέτρου α, τόσο πιο απότομη είναι η ευθεία γραμμή στη γραφική παράσταση.

5.3.1.2 Εκθετική συνάρτηση

Για να αναγνωριστεί ότι ένας πίνακας αποτελούμενος από τιμές x και y προέρχονται από μία εκθετική συνάρτηση $y = y_0 \times \alpha^x$, γίνεται εξέταση των αναλογιών των τιμών y, οι οποίες είναι σταθερές για τιμές x ίσης διαφοράς. Η γραφική παράσταση μίας συνάρτησης είναι κοίλη προς τα πάνω, ή κυρτή, εάν λυγίζει προς τα πάνω όσο γίνεται μετακίνηση από αριστερά προς τα δεξιά, ενώ η γραφική παράσταση μίας συνάρτησης είναι κοίλη προς τα κάτω, ή κοίλη, εάν λυγίζει προς τα κάτω όσο γίνεται μετακίνηση από τα αριστερά προς τα δεξιά. Μία ευθεία γραμμή δεν είναι ούτε κοίλη, ούτε κυρτή.

Μία συνάρτηση y είναι εκθετική συνάρτηση συναρτήσει του x με βάση α, εάν ισχύει $y = y_0 \times \alpha^x$ όπου το y_0 είναι μία αρχική ποσότητα και α είναι ο παράγοντας με τον οποίο το y μεταβάλλεται όταν το x αυξάνεται κατά μία μονάδα. Για τιμές α, οι οποίες βρίσκονται στο διάστημα $(1, +\infty)$ υπάρχει εκθετική αύξηση, ενώ για τιμές α οι οποίες βρίσκονται στο διάστημα $(0,1)$ υπάρχει εκθετική μείωση. Δεδομένου ότι $\alpha > 0$, το μεγαλύτερο δυνατό πεδίο τιμών για την εκθετική συνάρτηση είναι όλοι οι πραγματικοί αριθμοί. Ο λόγος για τον οποίο οι τιμές δεν μπορούν να ανήκουν στο διάστημα $(-\infty, 0]$ είναι, για παράδειγμα, δεν μπορεί να οριστεί $\alpha^{\frac{1}{4}}$ εάν $\alpha < 0$. Επίσης, η τιμή $\alpha = 1$ δεν είναι συχνή, αφού $y = y_0 \times 1^x$ είναι μία σταθερή συνάρτηση. Η τιμή του α είναι στενά συνδεδεμένη με το ποσοστό ρυθμού αύξησης ή μείωσης. Για παράδειγμα, εάν $\alpha = 1{,}04$ τότε το y αυξάνεται κατά 4%, εάν $\alpha = 0{,}96$, τότε το y μειώνεται κατά 4%.

Ο χρόνος μισής ζωής μίας ποσότητας, η οποία μειώνεται εκθετικά, είναι ο χρόνος που απαιτείται ώστε η ποσότητα να μειωθεί κατά ένα παράγοντα ενός δευτέρου. Ο χρόνος διπλασιασμού μίας ποσότητας, η οποία αυξάνεται εκθετικά, είναι ο χρόνος που απαιτείται έτσι ώστε η ποσότητα να διπλασιαστεί.

Η $y = y_0 \times \alpha^x$ δίνει μία οικογένεια από εκθετικές συναρτήσεις με θετικές παραμέτρους y_0, η οποία είναι η αρχική ποσότητα, και α, η οποία είναι η βάση ή ο παράγοντας αύξησης ή μείωσης. Η βάση καθορίζει το εάν η συνάρτηση αυξάνεται, για τιμές $(1, +\infty)$, ή μειώνεται, για τιμές $(0,1)$. Επειδή το α είναι ο παράγοντας με τον οποίο το y αλλάζει όταν το x αυξάνεται κατά ένα, μεγάλες τιμές του α εκφράζουν ταχεία αύξηση, ενώ τιμές του α οι οποίες είναι κοντά στη μηδενική εκφράζουν ταχεία μείωση. Όλα τα μέλη της οικογένειας $y = y_0 \times \alpha^x$ έχουν γραφική παράσταση, η οποία είναι κοίλη προς τα πάνω ή κοίλη.

Η πιο συχνά χρησιμοποιούμενη βάση για μία εκθετική συνάρτηση είναι ο γνωστός αριθμός e, ο οποίος έχει τιμή $2{,}7182$ .... Ο αριθμός αυτός χρησιμοποιείται συχνά ως βάση, αφού πολλοί τύποι υπολογισμού λύνονται πιο εύκολα όταν το e χρησιμοποιείται ως βάση. Μία εκθετικά αυξανόμενη συνάρτηση μπορεί να γραφεί για κάποιο α που ανήκει στο διάστημα $(1, +\infty)$ και κάποιο β που ανήκει στο διάστημα $(0, +\infty)$ στη μορφή:

$$y = y_0 \times \alpha^x \text{ ή } y = y_0 \times e^{\beta x}$$

και μία εκθετικά φθίνουσα συνάρτηση μπορεί να γραφεί για κάποιο α που ανήκει στο διάστημα $(0,1)$ και για κάποιο β που ανήκει στο διάστημα $(0, +\infty)$ στη μορφή:

$$y = y_0 \times \alpha^x \text{ ή } y = y_0 \times e^{-\beta x}$$

όπου y_0 είναι η αρχική ποσότητα. Το y μειώνεται ή αυξάνεται με ένα συνεχή ρυθμό β. Για παράδειγμα, τιμή $\beta = 0{,}04$ αντιστοιχεί σε ένα συνεχή ρυθμό 4%.

5.3.2 Αυξανόμενες και φθίνουσες συναρτήσεις

Οι όροι αυξανόμενη και φθίνουσα μπορούν να χαρακτηρίσουν τόσο τις γραμμικές συναρτήσεις, όσο και τις εκθετικές συναρτήσεις. Μία συνάρτηση f είναι αυξανόμενη, εάν οι τιμές $f(x)$ αυξάνονται όσο αυξάνεται το x, και μία συνάρτηση f είναι φθίνουσα εάν οι τιμές $f(x)$ μειώνονται όσο αυξάνεται το x. Η γραφική παράσταση μίας αυξανόμενης συνάρτησης έχει ανοδική τάση, όσο γίνεται μετακίνηση από τα αριστερά στα δεξιά, ενώ η γραφική παράσταση μίας φθίνουσας συνάρτησης έχει καθοδική τάση όσο γίνεται μετακίνηση από τα αριστερά στα δεξιά. Μία συνάρτηση $f(x)$ είναι μονοτονική, εάν αυξάνεται για όλες τις τιμές x ή μειώνεται για όλες τις τιμές x.

5.3.3 Αναλογικότητα

Μία συχνή συναρτησιακή σχέση είναι η σχέση αναλογικότητας, η οποία εμφανίζεται όταν ένα μέγεθος είναι ανάλογο προς ένα άλλο μέγεθος. Για παράδειγμα, το μέγεθος εμβαδού ενός τετραγώνου είναι ανάλογο του τετραγώνου του μήκους της πλευράς του. Μία συνάρτηση y είναι άμεσα ανάλογη με ένα μέγεθος x, εάν υπάρχει μία μη μηδενική σταθερά α, σύμφωνα με την οποία:

$$y = \alpha \times x$$

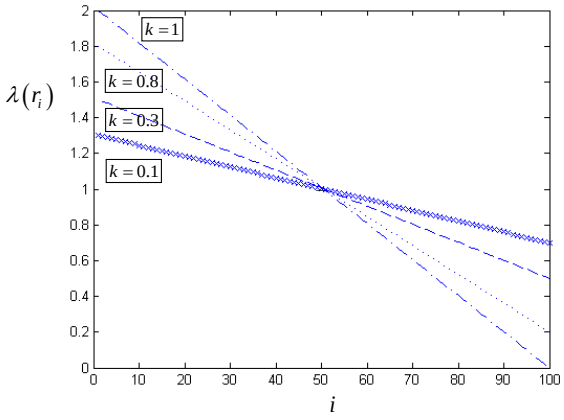
Η σταθερά α ονομάζεται σταθερά της αναλογικότητας.

5.4 Γραμμική και εκθετική συνάρτηση απόσβεσης

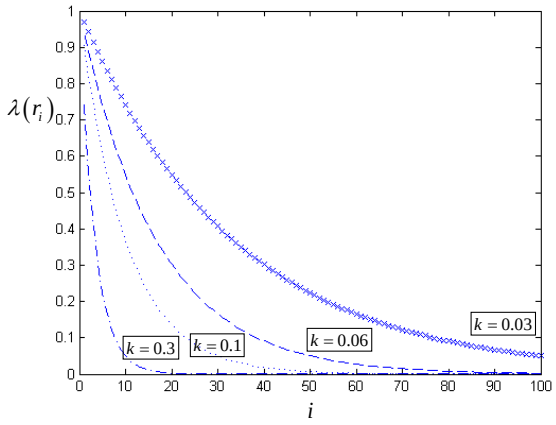
Παρουσιάζεται μία συνάρτηση απόσβεσης $\lambda(\kappa_t) = f(t)$ όπου κ_t είναι ένας κανόνας που προέκυψε πριν από t χρονικά βήματα. Σε κάθε χρονικό βήμα, κάθε κανόνας κ_t συνοδεύεται από μία τιμή βάρους, η οποία εξαρτάται από το πόσο παλιός είναι ο κανόνας. Το βάρος κάθε κανόνα αντιπροσωπεύει τη σημαντικότητα του κανόνα στο τρέχον χρονικό βήμα. Σε γενικές γραμμές, μπορούν να χρησιμοποιηθούν διαφορετικές $f(t)$, ανάλογα με

τη συγκεκριμένη περίπτωση, για την οποία γίνεται επεξεργασία των συμβάντων. Στην ενότητα αυτή γίνεται συζήτηση μίας γραμμικής και μίας εκθετικής προσέγγισης.

Ένα σύστημα, το οποίο χρησιμοποιεί μία γραμμική συνάρτηση απόσβεσης με $\lambda(\kappa_i) = -\frac{2k}{n-1}(i-1) + k + 1$ παρουσιάζεται στο [82] όπου n είναι ο αριθμός των προηγούμενων βημάτων που λαμβάνονται υπόψη, i είναι ένας μετρητής που ξεκινάει από το τρέχον χρονικό βήμα και πηγαίνει πίσω με την πάροδο του χρόνου, κ_i είναι ένας κανόνας ο οποίος προέκυψε πριν από i χρονικά βήματα και $k \in [0,1]$ είναι το ποσοστό μείωσης του βάρους ενός κανόνα σε κάθε χρονικό βήμα. Με την μεταβολή της τιμής k, μπορεί να ρυθμιστεί η κλίση της συνάρτησης απόσβεσης. Με παρόμοιο τρόπο, ένα σύστημα εκθετικής απόσβεσης με $\lambda(\kappa_i) = e^{-ki}$ παρουσιάζεται στο [83]. Η παράμετρος k καθορίζει πόσο γρήγορα μειώνονται τα βάρη με την πάροδο του χρόνου. Για μεγαλύτερες τιμές του k, γίνεται εκχώρηση μικρότερης τιμής βάρους στους κανόνες. Εάν ισχύει $k = 0$, τότε γίνεται εκχώρηση της ίδιας τιμής βάρους σε όλα τα χρονικά βήματα. Στις εικόνες 10 και 11 απεικονίζονται η γραμμική και εκθετική συνάρτηση απόσβεσης που περιεγράφηκαν προηγουμένως, παραμετροποιημένες με διακριτές τιμές k.



Εικόνα 10: Γραμμική συνάρτηση απόσβεσης, k=0,3, 0,5, 0,8, 1 και n=100



Εικόνα 11: Εκθετική συνάρτηση απόσβεσης με k=0,03, 0,06, 0,1, 0,3 και n=100

Εάν ένας κανόνας, ο οποίος αντιπροσωπεύει ακριβώς την ίδια εξάρτηση μεταξύ συμβάντων παράγεται σε πολλαπλά χρονικά βήματα, το προτεινόμενο πλαίσιο κάνει εξισορρόπηση μεταξύ των πιθανά μεταβαλλόμενων πιθανοτήτων των κανόνων, σταθμίζοντας τις αντίστοιχες τιμές πιθανοτήτων σύμφωνα με τις τιμές βάρους των κανόνων μέσα στο χρόνο. Συγκεκριμένα, γίνεται παρακολούθηση των κανόνων που προκύπτουν μέσα στα προηγούμενα $\Delta t_{\mu\nu\acute{\eta}\mu\eta}$ χρονικά βήματα και χρησιμοποιείται η ακόλουθη μορφή καθορισμού βαρών, για τον υπολογισμό της πιθανότητας ενός κανόνα, ο οποίος προέκυψε μέσα σε αυτό το χρονικό διάστημα:

$$P_{\kappa,\Delta t} = \frac{\sum_{i=t-\Delta t+1}^{t} \lambda(\kappa_i) \times P_{\kappa,i}}{\sum_{i=t-\Delta t+1}^{t} \lambda(\kappa_i)}$$

Για παράδειγμα, στην περίπτωση που $\Delta t_{\mu\nu\acute{\eta}\mu\eta} = 3$ και ένας κανόνας κ, ο οποίος προέκυψε πριν από δύο χρονικά βήματα με πιθανότητα $P_{\kappa,t-2} = 0{,}8$ και στο τρέχον χρονικό βήμα με πιθανότητα $P_{\kappa,t} = 0{,}3$. Εάν χρησιμοποιηθεί η γραμμική συνάρτηση απόσβεσης με $k = 0{,}8$, η τελική πιθανότητα που εκχωρείται στο κανόνα ισούται με $P_{\kappa,\Delta t} = \frac{(0{,}8\times 1)+(0{,}3\times 1{,}8)}{2{,}8} = 0{,}47$. Με αυτό το τρόπο, ένα σύστημα συσχέτισης χωρίς μνήμη μπορεί να αποκτήσει δυνατότητες μνήμης μέσα σε ένα χρονικό διάστημα $\Delta t_{\mu\nu\acute{\eta}\mu\eta}$.

6. ΥΛΟΠΟΙΗΣΗ ΣΥΣΤΗΜΑΤΟΣ

Στα πλαίσια της εργασίας έγινε υλοποίηση ενός συστήματος στη γλώσσα προγραμματισμού Java με βάση θεωρία και αλγορίθμους, οι οποίοι περιεγράφηκαν στις παραπάνω ενότητες. Από τη θεωρία που περιγράφει την ανίχνευση συμβάντων σε ροές δεδομένων αισθητήρων πολλαπλών μεταβλητών, χρησιμοποιήθηκαν ο αλγόριθμος συσσωρευτικού αθροίσματος και ο αλγόριθμος διαγραμμάτων ελέγχου Shewhart. Το συστατικό της υλοποίησης που υλοποιεί την ανίχνευση συμβάντων σε ροές δεδομένων αισθητήρων πολλαπλών μεταβλητών δέχεται ως είσοδο ένα διάνυσμα πλαισίου, και δίνει ως έξοδο ένα διάνυσμα συμβάντων, σε κάθε χρονικό βήμα.

Στη συνέχεια, από τη θεωρία που περιγράφει τη συσχέτιση συμβάντων σε ροές δεδομένων συμβάντων πολλαπλών μεταβλητών, χρησιμοποιήθηκε ο αλγόριθμος μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών, ο οποίος δέχεται ως είσοδο, σε κάθε χρονικό βήμα, ένα διάνυσμα συμβάντων που προέκυψε από την ανίχνευση συμβάντων σε ροές δεδομένων αισθητήρων πολλαπλών μεταβλητών. Τέλος, από τη θεωρία που περιγράφει το προσαρμοστικό φιλτράρισμα εξαρτήσεων συμβάντων, χρησιμοποιήθηκαν οι συναρτήσεις γραμμικής και εκθετικής απόσβεσης. Στη συνέχεια, παρουσιάζονται οι τεχνολογίες στις οποίες βασίζεται το σύστημα και μία σύντομη περιγραφή της υλοποίησης του συστήματος.

6.1 Αντικειμενοστραφής προγραμματισμός σε Java

Το πρότυπο αντικειμενοστραφούς προγραμματισμού έχει ως στόχο να παρέχει αρθρωτά και επαναχρησιμοποιούμενα προγράμματα και βιβλιοθήκες, μέσω της ενθυλάκωσης δεδομένων και κώδικα. Το μοντέλο αντικειμένων για την γλώσσα προγραμματισμού Java περιγράφεται στο [84]. Ένας τύπος συνδέεται με οντότητες του προγράμματος, όπως οι μεταβλητές και οι εκφράσεις. Ένας τύπος περιορίζει τις τιμές, τις οποίες μία οντότητα μπορεί να κατέχει, και καθορίζει τις λειτουργίες τις οποίες η οντότητα πρέπει να παρέχει. Μία κλάση ορίζει ένα νέο τύπο και περιγράφει πώς αυτός υλοποιείται. Ένα αντικείμενο είναι ένα στιγμιότυπο κλάσης ή ένας πίνακας. Κατά το χρόνο μεταγλώττισης, ένα αντικείμενο αναφέρεται ότι έχει ένα τύπο, ενώ κατά το χρόνο εκτέλεσης, ένα αντικείμενο αναφέρεται ότι ανήκει σε μία κλάση. Οι κλάσεις οι οποίες δηλώνονται ως αφηρημένες δεν μπορούν να αποκτήσουν υπόσταση, με άλλα λόγια δεν μπορούν να παραχθούν

αντικείμενα, τα οποία είναι στιγμιότυπα της κλάσης. Μία δήλωση διεπαφής καθορίζει ένα τύπο, ο οποίος αποτελείται από σταθερές και αφηρημένες μεθόδους.

Όλες οι κλάσεις έχουν μία άμεση υπερκλάση, με εξαίρεση την κλάση $Object$. Μία κλάση αναφέρεται ότι είναι άμεση υποκλάση των άμεσα υπερκλάσεων της. Μία άμεση υποκλάση αντλεί την υλοποίηση της από την άμεση υπερκλάση της. Οι κλάσεις, οι οποίες δηλώνονται ως τελικές, δεν μπορούν να έχουν υποκλάσεις. Η σχέση άντλησης υλοποιήσεων μεταξύ των κλάσεων διαμορφώνει μία ιεραρχία κλάσεων. Ένας τύπος T_1 είναι συμβατός ως προς την εκχώρηση με ένα τύπο T_2, σύμφωνα με κάποιους κανόνες. Εάν ο τύπος T_1 είναι τύπος κλάσης, και ο τύπος T_2 είναι τύπος κλάσης, τότε ο τύπος T_1 και ο τύπος T_2 πρέπει να είναι η ίδια κλάση, ή ο τύπος T_1 πρέπει να είναι υποκλάση του τύπου T_2. Εάν ο τύπος T_1 είναι τύπος κλάσης και ο τύπος T_2 είναι τύπος διεπαφής, τότε ο τύπος T_1 πρέπει να υλοποιεί το τύπο T_2. Εάν ο τύπος T_1 είναι ένας τύπος διεπαφής, και ο τύπος T_2 είναι τύπος κλάσης, τότε ο τύπος T_2 πρέπει να είναι η κλάση $Object$. Εάν ο τύπος T_1 είναι ένας τύπος διεπαφής, και ο τύπος T_2 είναι τύπος διεπαφής, τότε ο τύπος T_1 και ο τύπος T_2 πρέπει να είναι η ίδια διεπαφή, ή ο τύπος T_1 είναι υποδιεπαφή του τύπου T_2.

Μία μέθοδος είναι εκτελέσιμος κώδικας, ο οποίος μπορεί να καλεστεί και συσχετίζεται με μία κλάση. Μία μέθοδος κλάσης καλείται σε σχέση με το τύπο της κλάσης, ενώ μία μέθοδος στιγμιοτύπου καλείται σε σχέση με ένα στιγμιότυπο της κλάσης. Οι μέθοδοι καλούνται σε ένα σημείο κλήσης μεθόδου. Η μέθοδος, η οποία περιέχει το σημείο κλήσης ονομάζεται μέθοδος καλούντος και η μέθοδος, η οποία καλείται ονομάζεται μέθοδος καλούμενου. Οι μέθοδοι έχουν ένα σταθερό αριθμό από τυπικές παραμέτρους, κάθε μία εκ των οποίων έχει ένα τύπο, και μπορεί να επιστρέψει μία τιμή στο καλούντα. Η υπογραφή μίας μεθόδου αποτελείται από το όνομα της μεθόδου και τον αριθμό και τύπο των τυπικών παραμέτρων της μεθόδου. Ο κατασκευαστής είναι εκτελέσιμος κώδικας, ο οποίος αρχικοποιεί στιγμιότυπα κλάσεων. Σε αντίθεση με τις μεθόδους, οι κατασκευαστές δεν μπορούν να καλεστούν απευθείας. Οι εγγενείς μέθοδοι υλοποιούνται σε κώδικα, ο οποίος εξαρτάται από την πλατφόρμα.

Μία κλάση, η οποία υλοποιεί άμεσα μία διεπαφή, έχει υλοποίηση όλων των αφηρημένων μεθόδων οι οποίες ορίζονται στη διεπαφή. Είναι πιθανό μία κλάση να δηλώσει μία μέθοδο, με την ίδια υπογραφή μίας μεθόδου, η οποία δηλώνεται στην υπερκλάση αυτής. Εάν η μέθοδος είναι μία μέθοδος στιγμιοτύπου, η μέθοδος αναφέρεται ότι υπερισχύει της μεθόδου της υπερκλάσης. Εάν η μέθοδος είναι μέθοδος κλάσης, η μέθοδος αναφέρεται ότι κρύβει τη μέθοδο της υπερκλάσης. Μία κλάση κληρονομεί τις

μεθόδους της άμεσα υπερκλάσης της και άμεσα υπερδιεπαφών της, οι οποίοι δεν υπερισχύονται και δεν κρύβονται από κάποια δήλωση στην κλάση.

6.2 Η εικονική μηχανή Java

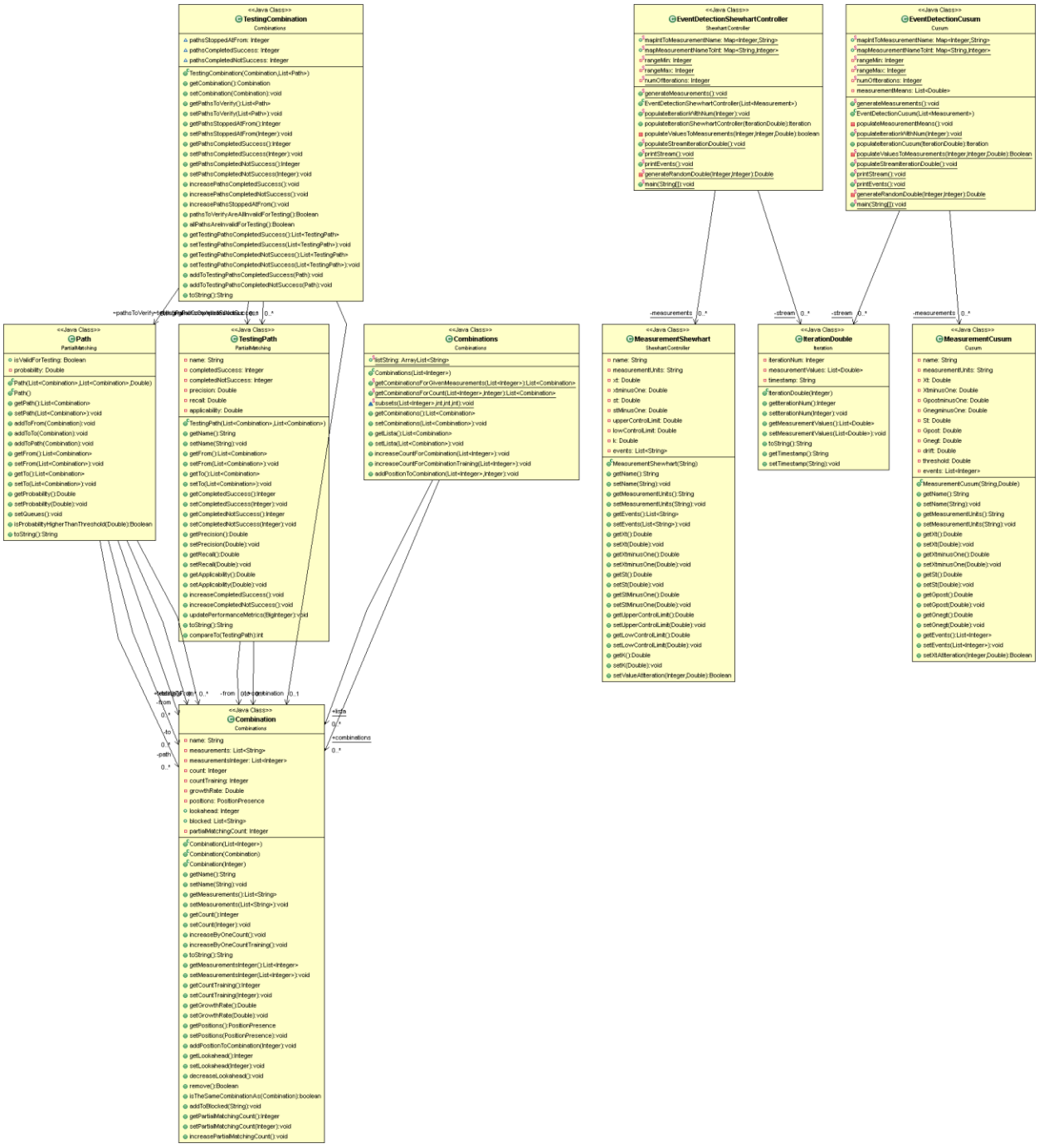
Η εικονική μηχανή Java [85] είναι μία αφηρημένη μηχανή, η οποία εκτελεί προγράμματα, τα οποία ορίζονται από τη μορφή αρχείου δυαδικής κλάσης. Το σύνολο εντολών της μορφής αρχείου κλάσης και της εικονικής μηχανής Java, το οποίο ονομάζεται κωδικοσύμβολα, υποστηρίζει όλες τις λειτουργίες, οι οποίες είναι απαραίτητες για την εκτέλεση προγραμμάτων, τα οποία είναι γραμμένα στη γλώσσα προγραμματισμού Java. Η εικονική μηχανή Java περιλαμβάνει ένα δυναμικό σωρό, μία περιοχή μεθόδων, ένα χώρο σταθερών χρόνου εκτέλεσης και δομές δεδομένων ανά νήμα. Η σωρός είναι ένας χώρος μνήμης, ο οποίος είναι κοινός για όλες τα νήματα, και στον οποίο αποθηκεύονται όλα τα αντικείμενα και πίνακες. Ο αυτόματος διαχειριστής μνήμης έχει την ευθύνη να απελευθερώνει τη μνήμη στη σωρό. Η περιοχή μεθόδων περιέχει δομές δεδομένων ανά κλάση, όπως η περιοχή σταθερών χρόνου εκτέλεσης και ο κώδικας μεθόδων και κατασκευαστών. Ο χώρος σταθερών χρόνου εκτέλεσης περιέχει σταθερές, οι οποίες αντιπροσωπεύουν στοιχεία του προγράμματος.

Κατά τη διάρκεια της εκτέλεσης, κάθε φορά που γίνεται κλήση κάποιας μεθόδου, δημιουργείται ένα πλαίσιο. Κάθε πλαίσιο περιέχει μία λίστα από τοπικές μεταβλητές, μία στοίβα τελεστών και ένα δείκτη στο χώρο σταθερών χρόνου εκτέλεσης της κλάσης, η οποία περιέχει τη μέθοδο που καλέστηκε. Οι τοπικές μεταβλητές διατηρούν τιμές, οι οποίες χρησιμοποιούνται κατά την εκτέλεση της μεθόδου. Η στοίβα τελεστών χρησιμοποιείται για τη διατήρηση μερικών αποτελεσμάτων και το πέρασμα παραμέτρων σε μεθόδους.

Το σύνολο εντολών της εικονικής μηχανής Java λειτουργεί κυρίως με τιμές, τις οποίες περιέχουν τοπικές μεταβλητές ή η στοίβα τελεστών. Το σύνολο εντολών περιλαμβάνει εντολές για εκτέλεση βασικών αριθμητικών πράξεων όπως πρόσθεση, πολλαπλασιασμός, λογική σύζευξη και συγκρίσεις. Επίσης, περιλαμβάνει εντολές για δημιουργία αντικειμένων, εκμετάλλευση της στοίβας τελεστών, μεταφορά ελέγχου, κλήση μεθόδων και συγχρονισμό νημάτων.

6.3 Περιγραφή υλοποίησης

Στη συνέχεια παρουσιάζεται το διάγραμμα ενός υποσυνόλου των βασικών κλάσεων του συστήματος. Για εξοικονόμηση χώρου, δεν γίνεται αναπαράσταση όλων των κλάσεων του συστήματος.



Εικόνα 12: Διάγραμμα κλάσεων ενός υποσυνόλου των βασικών κλάσεων της υλοποίησης συστήματος διαχείρισης συμβάντων σε πολλαπλών μεταβλητών δεδομένα αισθητήρων ροής

Στη συνέχεια γίνεται μία σύντομη περιγραφή των βασικότερων κλάσεων του συστήματος.

- **Κλάση Συνδυασμός**

Τα στιγμιότυπα της κλάσης $Combination$ αντιπροσωπεύουν μία μέτρηση ή ένα συνδυασμό μετρήσεων. Έχουν ως χαρακτηριστικά ένα όνομα μέτρησης ή συνδυασμού, το οποίο χρησιμοποιείται για ταυτοποίηση όμοιων μετρήσεων ή συνδυασμών, και ένα μετρητή, ο οποίος χρησιμοποιείται για τον υπολογισμό της πιθανότητας στον αλγόριθμο μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών. Ένα στιγμιότυπο της κλάσης $Combination$ αποτελεί ένα κόμβο σε ένα δένδρο στην ιστορική δομή του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών.

- **Κλάση Συνδυασμοί**

Η κλάση $Combinations$ περιέχει μεθόδους, οι οποίοι χρησιμοποιούνται για εύρεση πιθανών συνδυασμών μετρήσεων, δοσμένων κάποιων μετρήσεων.

- **Κλάση Συνδυασμός Ελέγχου**

Σε κάθε επανάληψη, μετά την επανάληψη που καθορίζεται από την παράμετρο εισόδου $StartTestingIteration$, για κάθε μέτρηση και συνδυασμό μετρήσεων για τις οποίες υπήρξε εμφάνιση συμβάντος, σύμφωνα με το διάνυσμα συμβάντων, δημιουργείται ένα στιγμιότυπο της κλάσης $TestingCombination$. Το στιγμιότυπο αυτό περιέχει το όνομα της μέτρησης ή συνδυασμού για λόγους ταυτοποίησης, μετρικές οι οποίες χρησιμοποιούνται για υπολογισμό των μετρικών αποτελεσματικότητας και ένα πίνακα από κανόνες προς έλεγχο. Οι κανόνες αυτοί προκύπτουν από την αποδόμηση του δένδρου στην ιστορική δομή του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών, το οποίο έχει ρίζα με όνομα μέτρησης ή συνδυασμού μετρήσεων όμοιο με το όνομα του στιγμιότυπου. Κάθε στιγμιότυπο έχει χρόνο ζωής όσο για τον έλεγχο των επιμέρους κανόνων, και χρησιμοποιείται αποκλειστικά κατά τη διαδικασία ελέγχου.

- **Κλάση Ανίχνευση Συμβάντων με τον αλγόριθμο συσσωρευτικού αθροίσματος**

Η κλάση $EventDetectionCusum$ χρησιμοποιείται ως βοηθητική για την ανίχνευση συμβάντων με τον αλγόριθμο συσσωρευτικού αθροίσματος.

- **Κλάση Μέτρηση αλγορίθμου συσσωρευτικού αθροίσματος**

Για κάθε μέτρηση στο αρχείο εισόδου, δημιουργείται ένα στιγμιότυπο της κλάσης $MeasurementCusum$, το οποίο χρησιμοποιείται για την ανίχνευση συμβάντων με τον αλγόριθμο συσσωρευτικού αθροίσματος για τη συγκεκριμένη μέτρηση. Η κλάση $MeasurementCusum$ περιλαμβάνει τις παραμέτρους του αλγορίθμου, όπως για παράδειγμα μέσο όρο, ανώτατο όριο ελέγχου και κατώτατο όριο ελέγχου, και σε κάθε επανάληψη, δηλαδή σε κάθε γραμμή του αρχείου εισόδου, καλείται η μέθοδος $setXAtIteration$, η οποία εκτελεί τον αλγόριθμο και καθορίζει την ύπαρξη συμβάντος.

- **Αρχείο καταμέτρησης Αλγόριθμος Ανίχνευσης Συμβάντων**

Το αρχείο καταμέτρησης $EventDetectionAlgorithmEnum$ χρησιμοποιείται για την επιλογή του αλγορίθμου, ο οποίος χρησιμοποιείται στη διαδικασία ανίχνευσης συμβάντων.

- **Κλάση Γενικό Δένδρο**

Κάθε στιγμιότυπο της κλάσης $GenericTree$ αποτελεί ένα δένδρο, το οποίο είναι συστατικό της ιστορικής δομής του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών. Η κλάση $GenericTree$ περιλαμβάνει όλες τις

απαραίτητες μεθόδους για την δημιουργία, επεξεργασία και ανάκτηση γνώσης από το δένδρο, όπως η διάσχιση του δένδρου, καθορισμός και ανάκτηση ρίζας και άλλα.

- **Κλάση Κόμβος Γενικού Δένδρου**

Κάθε στιγμιότυπο της κλάσης $GenericTreeNode$ αποτελεί ένα κόμβο κάποιου στιγμιότυπου της κλάσης $GenericTree$. Η κλάση $GenericTreeNode$ περιλαμβάνει τα δεδομένα του κόμβου, δηλαδή ένα στιγμιότυπο της κλάσης $Combination$, μία λίστα από δείκτες στα άμεσα παιδιά του κόμβου, ένας δείκτης στο γονικό κόμβο και ένας αριθμός, ο οποίος καθορίζει το επίπεδο του κόμβου. Το επίπεδο του κόμβου είναι μηδενικό, όταν ο κόμβος είναι ρίζα του δένδρου, και αυξάνεται κατά ένα σε κάθε επόμενο επίπεδο του δένδρου. Επίσης, η κλάση περιέχει μεθόδους, οι οποίες βοηθούν στη δημιουργία ενός δένδρου και στην εκτέλεση του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών, όπως προσθήκη κόμβου - παιδί, αφαίρεση κόμβου - παιδί, ανάκτηση δεδομένων κόμβου και εύρεση συνδυασμού.

- **Αρχείο Καταμέτρησης Διάταξη Διάσχισης Γενικού Δένδρου**

Το αρχείο καταμέτρησης $GenericTreeTraversalOrderEnum$ περιλαμβάνει τις πιθανές επιλογές διάσχισης ενός δένδρου στην ιστορική δομή του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών. Οι πιθανές επιλογές είναι η διάσχιση του δένδρου με προδιάταξη και η διάσχιση του δένδρου με μεταδιάταξη.

- **Κλάση Επανάληψη**

Η κλάση $Iteration$ αντιπροσωπεύει ένα διάνυσμα συμβάντων, το οποίο αποτελεί έξοδος από τον αλγόριθμο ανίχνευσης συμβάντων και είσοδος στον αλγόριθμο μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών. Κάθε στιγμιότυπο της κλάσης $Iteration$ έχει χρόνο ζωής ίσο με το χρόνο επεξεργασίας μίας γραμμής του αρχείου εισόδου και περιλαμβάνει ένα ακέραιο αριθμό, ο οποίος καθορίζει τον τρέχον

αριθμό γραμμής του αρχείου εισόδου και ένα δυαδικό πίνακα, που καθορίζει την ύπαρξη συμβάντος για κάθε μία από τις μετρήσεις. Περιλαμβάνει επίσης μεθόδους για υποβοήθηση του αλγορίθμου μεταβλητής τάξης συσχέτισης συμβάντων πολλαπλών μεταβλητών, όπως για παράδειγμα την ανάκτηση των μετρήσεων, στις οποίες εμφανίστηκε συμβάν.

- **Κλάση Επανάληψη Πραγματικών Αριθμών**

Η κλάση $IterationDouble$ αντιπροσωπεύει ένα διάνυσμα πλαισίου, το οποίο αποτελεί είσοδος στον αλγόριθμο ανίχνευσης συμβάντων και προκύπτει από μία γραμμή του αρχείου εισόδου. Κάθε στιγμιότυπο της κλάσης $IterationDouble$ έχει χρόνο ζωής ίσο με το χρόνο επεξεργασίας μίας γραμμής του αρχείου εισόδου, και περιλαμβάνει ένα ακέραιο αριθμό, ο οποίος καθορίζει τον τρέχον αριθμό γραμμής του αρχείου εισόδου, και ένα πίνακα από πραγματικούς αριθμούς, όπου κάθε πραγματικός αριθμός είναι η τιμή μίας μέτρησης στο συγκεκριμένο χρονικό βήμα.

- **Κλάση Επεξεργασία**

Η κλάση $Process$ περιλαμβάνει την κύρια μέθοδο της υλοποίησης, και όλες τις βασικές μεθόδους του συστήματος, όπως για παράδειγμα μέθοδοι για την εκτέλεση των αλγορίθμων ανίχνευσης συμβάντων, του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών, της διαδικασίας προσαρμοστικού φιλτραρίσματος και άλλα. Στη συνέχεια παρουσιάζεται η αντιπροσώπευση της κλάσης $Process$, η οποία περιλαμβάνει όλα τα χαρακτηριστικά και τις μεθόδους της κλάσης.

Εικόνα 13: Κλάση Επεξεργασία της υλοποίησης συστήματος διαχείρισης συμβάντων σε πολλαπλών μεταβλητών δεδομένα αισθητήρων ροής

- **Αρχείο καταμέτρησης Συνάρτηση Απόσβεσης**

Το αρχείο καταμέτρησης $AgingFunctionEnum$ περιλαμβάνει τις διαθέσιμες επιλογές της διαδικασίας προσαρμοστικού φιλτραρίσματος. Υπάρχει η επιλογή της μη εκτέλεσης κάποιας διαδικασίας προσαρμοστικού φιλτραρίσματος, ή της εκτέλεσης της διαδικασίας προσαρμοστικού φιλτραρίσματος χρησιμοποιώντας τη γραμμική συνάρτηση απόσβεσης ή την εκθετική συνάρτηση απόσβεσης.

- **Κλάση Μονοπάτι**

Η κλάση $Path$ αντιπροσωπεύει ένα χρονικό κανόνα συσχέτισης, ο οποίος προέκυψε από ένα μονοπάτι ενός δένδρου της ιστορικής δομής του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών. Χρησιμοποιείται αποκλειστικά κατά τη διαδικασία εκπαίδευσης μοντέλου του αλγορίθμου, και χρησιμοποιείται επίσης για την δημιουργία χρονικών κανόνων συσχέτισης της διαδικασίας ελέγχου. Περιλαμβάνει ένα πίνακα από στιγμιότυπα της κλάσης $Combination$, κάποιους πίνακες που αποτελούν υποσύνολα του πρώτου πίνακα, και μία πραγματική τιμή πιθανότητας του κανόνα. Ο πίνακας, ο οποίος αποτελείται από στιγμιότυπα της κλάσης $Combination$ αντιπροσωπεύει ένα μονοπάτι σε ένα δένδρο του αλγορίθμου, το οποίο αποτελείται από κόμβους του δένδρου. Οι πίνακες, οι οποίοι περιλαμβάνουν υποσύνολα του πρώτου πίνακα χρησιμοποιούνται για την υποβοήθηση των διαδικασιών, στις οποίες είναι απαραίτητος ο διαχωρισμός κεφαλής από το σώμα του κανόνα, διαχωρισμός ο οποίος προκύπτει σύμφωνα με τις τιμές παραμέτρων m και l του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών.

- **Κλάση Πιθανότητες Μονοπατιού**

Κάθε στιγμιότυπο της κλάσης $ProbabilityPath$ προκύπτει από ένα στιγμιότυπο της κλάσης $Path$ και αντιπροσωπεύει ένα χρονικό κανόνα συσχέτισης. Κάθε στιγμιότυπο της κλάσης $ProbabilityPath$ περιλαμβάνει μία μεταβλητή, η οποία περιλαμβάνει την

ονοματολογία του κανόνα και χρησιμοποιείται για σκοπούς ταυτοποίησης κανόνων, και ένα πίνακα από πραγματικές τιμές πιθανοτήτων του κανόνα για τα n προηγούμενα χρονικά βήματα. Τα στιγμιότυπα αυτά δημιουργούνται μόνο εάν επιλεγεί η εκτέλεση της διαδικασίας προσαρμοστικού φιλτραρίσματος και χρησιμοποιείται αποκλειστικά από τη διαδικασία αυτή.

- **Κλάση Μονοπάτι Ελέγχου**

Κάθε στιγμιότυπο της κλάσης $TestingPath$ προκύπτει από ένα στιγμιότυπο της κλάσης $Path$ και αντιπροσωπεύει ένα χρονικό κανόνα συσχέτισης. Τα στιγμιότυπα της κλάσης $TestingPath$ χρησιμοποιούνται αποκλειστικά από τη διαδικασία ελέγχου της υλοποίησης, και αποθηκεύονται σε μία δομή της κλάσης που περιλαμβάνει την κύρια μέθοδο του προγράμματος $Process$. Τα στιγμιότυπα περιλαμβάνουν μία μεταβλητή ονοματολογίας για ταυτοποίηση κανόνων, όπως επίσης και πίνακες από στιγμιότυπα της κλάσης $Combination$ που αντιπροσωπεύουν υποσύνολα του μονοπατιού του δένδρου για καθορισμό της κεφαλής και σώματος του κανόνα. Επίσης περιλαμβάνει μετρητές ακέραιων αριθμών $ΟλοκληρώθηκεΕπιτυχώς$ και $ΟλοκληρώθηκεΜηΕπιτυχώς$, οι οποίοι χρησιμοποιούνται για τον υπολογισμό των μετρικών αποτελεσματικότητας του συστήματος, αλλά και κάθε χρονικού κανόνα συσχέτισης.

- **Κλάση Ανίχνευση Συμβάντων με τον αλγόριθμο διαγραμμάτων ελέγχου Shewhart**

Η κλάση $EventDetectionShewhart$ χρησιμοποιείται ως βοηθητική για την ανίχνευση συμβάντων με τον αλγόριθμο διαγραμμάτων ελέγχου Shewhart.

- **Κλάση Μέτρηση αλγορίθμου διαγραμμάτων ελέγχου Shewhart**

Για κάθε μέτρηση στο αρχείο εισόδου, δημιουργείται ένα στιγμιότυπο της κλάσης $MeasurementShewhart$, το οποίο χρησιμοποιείται για την ανίχνευση συμβάντων με τον αλγόριθμο διαγραμμάτων ελέγχου Shewhart για τη συγκεκριμένη μέτρηση. Η κλάση

$MeasurementShewhart$ περιλαμβάνει τις παραμέτρους του αλγορίθμου, όπως για παράδειγμα μέσο όρο, τυπική απόκλιση, ανώτατο όριο ελέγχου και κατώτατο όριο ελέγχου και σε κάθε επανάληψη, δηλαδή σε κάθε γραμμή του αρχείου εισόδου, καλείται η μέθοδος $setXAtIteration$, η οποία εκτελεί τον αλγόριθμο και καθορίζει την ύπαρξη συμβάντος.

7. ΠΕΙΡΑΜΑΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ

Τα πειράματα που διεξήχθησαν στα πλαίσια της εργασίας βασίζονται στη θεωρία, η οποία περιεγράφηκε στις προηγούμενες ενότητες. Ως είσοδος χρησιμοποιήθηκε ένα αρχείο διαχώρισης τιμών με κόμμα, το οποίο περιέχει καταγραφές 29 μετρήσεων κάθε πέντε λεπτά. Το πρόγραμμα έχει υλοποιηθεί με τέτοιο τρόπο, ώστε να μπορεί να χαρακτηριστεί ως πρόγραμμα σε πραγματικό χρόνο ή κατά προσέγγιση σε πραγματικό χρόνο. Με άλλα λόγια, η υλοποίηση στα πλαίσια της εργασίας αυτής μπορεί να χρησιμοποιηθεί σε δεδομένα συνεχούς ροής, με κάποιες απλές αλλαγές στο πηγαίο κώδικα. Επίσης, έχουν ληφθεί υπόψη κάποιες επιπλέον παράμετροι, οι οποίοι περιγράφονται στη συνέχεια της ενότητας αυτής, έτσι ώστε να επιτυγχάνεται εξοικονόμηση των πόρων του συστήματος στο οποίο γίνεται η εκτέλεση, και ταυτοχρόνως η εκτέλεση να γίνεται με μεγαλύτερη ταχύτητα. Οι πόροι αυτοί είναι υπολογιστικοί πόροι και πόροι μνήμης του συστήματος. Στη συνέχεια της ενότητας αυτής γίνεται περιγραφή των δεδομένων και αλγορίθμων που χρησιμοποιήθηκαν. Επίσης, γίνεται περιγραφή των πειραμάτων που εκτελέστηκαν, και μελέτη των συμπερασμάτων που εξήχθησαν από την εκτέλεση των πειραμάτων αυτών.

7.1 Δεδομένα

Τα δεδομένα, τα οποία χρησιμοποιήθηκαν στα πλαίσια της εργασίας προέρχονται από το τομέα της ναυτιλίας. Συγκεκριμένα, τα δεδομένα προέρχονται από ένα δίκτυο αισθητήρων, το οποίο είναι κατανεμημένα τοποθετημένο σε ένα πλοίο. Τα δεοδμένα έχουν τη μορφή χρονοσειρών, οι οποίες έχουν χρονική διαφορά πέντε λεπτών μεταξύ τους. Κάθε μέτρηση αποτελεί μία μεταβλητή, και το σύστημα αποτελείται συνολικά από 29 μεταβλητές, εκ των οποίων η μία μεταβλητή αποτελεί το χρονικό προσδιορισμό της χρονοσειράς, δηλαδή ημερομηνία και ώρα στην οποία έγιναν οι μετρήσεις. Να σημειωθεί ότι η διάσταση χρονικού προσδιορισμού δεν χρησιμοποιήθηκε για την επεξεργασία. Οι μεταβλητές, σε ένα μεγάλο βαθμό, αφορούν το μηχανικό σκέλος του πλοίου. Παραδείγματα μεταβλητών είναι η ροπή, το βάθος, η ταχύτητα του αέρα και άλλα. Τα δεδομένα αισθητήρων αποθηκεύτηκαν σε ένα αρχείο διαχώρισης τιμών με κόμμα το οποίο έχει 29 στήλες και 21145 γραμμές, όπου κάθε γραμμή αποτελεί μία χρονοσειρά μετρήσεων κάποιας χρονικής στιγμής, και κάθε χρονοσειρά είναι επόμενη της χρονοσειράς που καταμετρήθηκε πριν από πέντε λεπτά, με άλλα λόγια οι χρονοσειρές είναι διατεταγμένες με βάση το χρόνο.

7.2 Περιγραφή διαδικασίας ελέγχου

Η διαδικασία ελέγχου της υλοποίησης ξεκινάει μετά από κάποιο χρονικό βήμα, ή γραμμή του αρχείου εισόδου, το οποίο καθορίζεται στην αρχή του κάθε πειράματος. Τα πειράματα που εκτελέστηκαν στα πλαίσια της εργασίας ξεκινούν τη διαδικασία ελέγχου στο βήμα 100. Όταν ξεκινάει η διαδικασία ελέγχου, για κάθε μέτρηση ή συνδυασμό μετρήσεων που δηλώνει συμβάν ή συνδυασμό συμβάντων στο κάθε χρονικό βήμα, δημιουργείται ένα στιγμιότυπο της κλάσης $Συνδυασμός Ελέγχου$. Κάθε αντικείμενο της κλάσης αυτής περιέχει ένα πίνακα από κανόνες προς επαλήθευση, οι οποίοι προκύπτουν από την αποδόμηση ενός δένδρου της ιστορικής δομής του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών, το οποίο έχει ως ρίζα το προς διερεύνηση συμβάν ή συνδυασμό συμβάντων.

Σε κάθε επόμενο χρονικό βήμα, ανάλογα με την εμφάνιση ή όχι κάποιου συμβάντος ή συνδυασμού συμβάντων, εκτελείται ένα από τα ακόλουθα για κάθε κανόνα:

- Η μέχρι τώρα επαλήθευση του κανόνα.
- Ακύρωση του κανόνα.
- Χαρακτηρισμός του κανόνα ως $Ολοκληρώθηκε Επιτυχώς$.
- Χαρακτηρισμός του κανόνα ως $Ολοκληρώθηκε Μη Επιτυχώς$.

Η μέχρι τώρα επαλήθευση δηλώνει ότι ο κανόνας επιβεβαιώνεται στο επόμενο χρονικό βήμα, σύμφωνα με τη δομή του κανόνα και την παρουσία συμβάντων ή συνδυασμού συμβάντων στο προς διερεύνηση χρονικό βήμα. Η ακύρωση κανόνα δηλώνει ότι ο κανόνας δεν επαληθεύεται όσο αφορά το σώμα του κανόνα για τιμές της παραμέτρου του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών $m > 1$, και συνεπώς χαρακτηρίζεται ως $Διακοπή Στο Σώμα$. Ο χαρακτηρισμός $Ολοκληρώθηκε Επιτυχώς$ ενός κανόνα δηλώνει ότι ο κανόνας έχει επαληθευθεί τόσο στο σώμα όσο και στην κεφαλή του κανόνα, και στο τρέχον χρονικό βήμα έχει επίσης επαληθευθεί, και δεν υπάρχει κάποιο άλλο στοιχείο του κανόνα που πρέπει να επαληθευθεί. Ο χαρακτηρισμός $Ολοκληρώθηκε Μη Επιτυχώς$ ενός κανόνα δηλώνει ότι ο κανόνας έχει επαληθευθεί στο σώμα του κανόνα, αλλά στην κεφαλή του κανόνα, για κάποιο στοιχείο, δεν υπάρχει επαλήθευση. Η μετρική ακρίβειας, που χρησιμοποιήθηκε στα πλαίσια των πειραμάτων, βασίζεται σε αυτές τις μετρικές.

7.3 Μετρικές αξιολόγησης πειραμάτων

Γενικά, οι μετρικές αξιολόγησης σε προβλήματα κατηγοριοποίησης ορίζονται από ένα πίνακα, με τους αριθμούς των παραδειγμάτων τα οποία κατηγοριοποιήθηκαν ορθά ή λανθασμένα για κάθε κλάση, ο οποίος ονομάζεται πίνακας συγχύσεως. Ο πίνακας συγχύσεως για ένα δυαδικό πρόβλημα κατηγοριοποίησης, δηλαδή ένα πρόβλημα κατηγοριοποίησης το οποίο έχει δύο μόνο κλάσεις $\theta\varepsilon\tau\iota\kappa\acute{o}$ ή $\alpha\rho\nu\eta\tau\iota\kappa\acute{o}$, φαίνεται στο πίνακα:

Πίνακας 2: Πίνακας Συγχύσεως

	Κλάση Πρόβλεψης	
Πραγματική Κλάση	**Θετική**	**Αρνητική**
Θετική	Αληθώς Θετικά	Ψευδώς Αρνητικά
Αρνητική	Ψευδώς Θετικά	Αληθώς Αρνητικά

Οι έννοιες $\Psi\varepsilon\upsilon\delta\acute{\omega}\varsigma\ \Theta\varepsilon\tau\iota\kappa\acute{\alpha}$, $\Psi\varepsilon\upsilon\delta\acute{\omega}\varsigma\ \Alpha\rho\nu\eta\tau\iota\kappa\acute{\alpha}$, $\Alpha\lambda\eta\theta\acute{\omega}\varsigma\ \Theta\varepsilon\tau\iota\kappa\acute{\alpha}$ και $\Alpha\lambda\eta\theta\acute{\omega}\varsigma\ \Alpha\rho\nu\eta\tau\iota\kappa\acute{\alpha}$ χρησιμοποιούνται ευρέως στην αξιολόγηση ενός συστήματος. Τα $\Psi\varepsilon\upsilon\delta\acute{\omega}\varsigma\ \Theta\varepsilon\tau\iota\kappa\acute{\alpha}$ ορίζονται ως τα παραδείγματα, τα οποία προβλέφθηκαν ως θετικά και προέρχονται από την αρνητική κλάση. Τα $\Psi\varepsilon\upsilon\delta\acute{\omega}\varsigma\ \Alpha\rho\nu\eta\tau\iota\kappa\acute{\alpha}$ είναι τα παραδείγματα, τα οποία προβλέφθηκαν ως αρνητικά, των οποίων η πραγματική κλάση είναι η θετική κλάση. Τα $\Alpha\lambda\eta\theta\acute{\omega}\varsigma\ \Theta\varepsilon\tau\iota\kappa\acute{\alpha}$ είναι τα παραδείγματα τα οποία προβλέφθηκαν ορθά ότι ανήκουν στη θετική κλάση. Τέλος, τα $\Alpha\lambda\eta\theta\acute{\omega}\varsigma\ \Alpha\rho\nu\eta\tau\iota\kappa\acute{\alpha}$ είναι τα παραδείγματα, τα οποία προβλέφθηκαν ορθά ότι ανήκουν στην αρνητική κλάση.

Η μετρική αξιολόγησης, η οποία χρησιμοποιείται ευρέως στην πράξη είναι το ποσοστό ακρίβειας. Η μετρική αυτή αξιολογεί την αποτελεσματικότητα ενός κατηγοριοποιητή από το ποσοστό των ορθών προβλέψεων. Το ποσοστό ακρίβειας ορίζεται ως:

$$\Alpha\kappa\rho\acute{\iota}\beta\varepsilon\iota\alpha = \frac{|\Alpha\lambda\eta\theta\acute{\omega}\varsigma\ \Alpha\rho\nu\eta\tau\iota\kappa\acute{\alpha}| + |\Alpha\lambda\eta\theta\acute{\omega}\varsigma\ \Theta\varepsilon\tau\iota\kappa\acute{\alpha}|}{|\Psi\varepsilon\upsilon\delta\acute{\omega}\varsigma\ \Alpha\rho\nu\eta\tau\iota\kappa\acute{\alpha}| + |\Psi\varepsilon\upsilon\delta\acute{\omega}\varsigma\ \Theta\varepsilon\tau\iota\kappa\acute{\alpha}| + |\Alpha\lambda\eta\theta\acute{\omega}\varsigma\ \Alpha\rho\nu\eta\tau\iota\kappa\acute{\alpha}| + |\Alpha\lambda\eta\theta\acute{\omega}\varsigma\ \Theta\varepsilon\tau\iota\kappa\acute{\alpha}|}$$

όπου $|X|$ συμβολίζει την πληθικότητα του συνόλου X. Η συμπληρωματική μετρική του ποσοστού ακρίβειας είναι το ποσοστό σφάλματος, η οποία αξιολογεί ένα κατηγοριοποιητή από το ποσοστό των εσφαλμένων προβλέψεων. Η μετρική ποσοστού σφάλματος ορίζεται ως:

$$Ποσοστό\ Σφάλματος = \frac{|Ψευδώς\ Αρνητικά| + |Ψευδώς\ Θετικά|}{|Ψευδώς\ Αρνητικά| + |Ψευδώς\ Θετικά| + |Αληθώς\ Αρνητικά| + |Αληθώς\ Θετικά|}$$

Οι μετρικές ποσοστού ακρίβειας και ποσοστού σφάλματος είναι γενικές μετρικές, και συνεπώς μπορούν να προσαρμοστούν άμεσα σε προβλήματα κατηγοριοποίησης πολλαπλών κλάσεων.

Οι μετρικές ανάκλησης και εξειδίκευσης αξιολογούν την αποτελεσματικότητα ενός κατηγοριοποιητή, για κάθε κλάση στο δυαδικό πρόβλημα. Η ανάκληση είναι γνωστή επίσης και ως ευαισθησία ή ρυθμός αληθών θετικών και είναι το ποσοστό των παραδειγμάτων, τα οποία ανήκουν στη θετική κλάση και προβλέφθηκαν ορθά ως θετικά. Η ανάκληση ορίζεται ως:

$$Ανάκληση = \frac{|Αληθώς\ Θετικά|}{|Αληθώς\ Θετικά| + |Ψευδώς\ Αρνητικά|}$$

Η εξειδίκευση είναι το ποσοστό των αρνητικών παραδειγμάτων, τα οποία προβλέφθηκαν ορθά ως αρνητικά. Η εξειδίκευση με μαθηματικούς ορισμούς ορίζεται ως:

$$Εξειδίκευση = \frac{|Αληθώς\ Αρνητικά|}{|Ψευδώς\ Θετικά| + |Αληθώς\ Αρνητικά|}$$

Η ακρίβεια είναι μία μετρική, η οποία προσεγγίζει την πιθανότητα μία θετική πρόβλεψη να είναι ορθή. Η ακρίβεια με μαθηματικούς τύπους ορίζεται ως:

$$Ακρίβεια = \frac{|Αληθώς\ Θετικά|}{|Αληθώς\ Θετικά| + |Ψευδώς\ Θετικά|}$$

Η μετρική ακρίβειας μπορεί να συνδυαστεί με τη μετρική ανάκλησης, χρησιμοποιώντας τη μετρική $F\ μέτρο$. Μία σταθερά β εξισορροπεί την επίδραση μεταξύ των μετρικών ακρίβειας και ανάκλησης. Η μετρική $F\ μέτρο$ με μαθηματικούς τύπους ορίζεται ως:

$$F\ μέτρο = \frac{(\beta^2 + 1) \times Ακρίβεια \times Ανάκληση}{\beta^2 \times Ακρίβεια + Ανάκληση}$$

Στο πλαίσιο του επιστημονικού τομέα της ανάκλησης πληροφορίας, η αποτελεσματικότητα ενός συστήματος ανάκλησης μπορεί να προσεγγιστεί με δύο πολύ συχνές και βασικές μετρικές που περιεγράφηκαν προηγουμένως, η μετρική ακρίβειας και

η μετρική ανάκλησης. Στην απλή περίπτωση ενός συστήματος ανάκτησης πληροφορίας που επιστρέφει ένα σύνολο εγγράφων για ένα ερώτημα χωρίς πληροφορία κατάταξης, οι μετρικές ακρίβειας και ανάκλησης έχουν παρόμοιο χαρακτήρα με ένα σύστημα πρόβλεψης.

Η ακρίβεια είναι το ποσοστό των εγγράφων που ανακτήθηκαν, τα οποία είναι σχετικά και δίνεται από:

$$Ακρίβεια = \frac{Αριθμός\ ανακτώμενων\ στοιχείων\ τα\ οποία\ είναι\ σχετικά}{Αριθμός\ ανακτώμενων\ στοιχείων}$$
$$= P(σχετικό\ στοιχείο|ανακτώμενο\ στοιχείο)$$

Η ανάκληση είναι το ποσοστό των σχετικών εγγράφων τα οποία ανακτήθηκαν. Η ανάκληση ορίζεται ως:

$$Ανάκληση = \frac{Αριθμός\ ανακτώμενων\ στοιχείων\ τα\ οποία\ είναι\ σχετικά}{Αριθμός\ σχετικών\ στοιχείων}$$
$$= P(ανακτώμενο\ στοιχείο|σχετικό\ στοιχείο)$$

Οι έννοιες της ακρίβειας και ανάκλησης μπορούν να είναι πιο σαφείς εξετάζοντας το πίνακα συγχύσεως στο πλαίσιο της ανάκτησης πληροφορίας:

Πίνακας 3: Πίνακας Συγχύσεως στο πλαίσιο της ανάκτησης πληροφορίας

	Σχετικό	**Μη Σχετικό**
Ανακτώμενο	Αληθώς Θετικά	Ψευδώς Θετικά
Μη Ανακτώμενο	Ψευδώς Αρνητικά	Αληθώς Αρνητικά

Τότε οι μετρικές ακρίβειας και ανάκλησης ορίζονται ως:

$$Ακρίβεια = \frac{Αληθώς\ Θετικά}{Αληθώς\ Θετικά + Ψευδώς\ Θετικά}$$

$$Ανάκληση = \frac{Αληθώς\ Θετικά}{Αληθώς\ Θετικά + Ψευδώς\ Αρνητικά}$$

Μπορεί να θεωρηθεί ότι σε ένα σύστημα ανάκτησης υπάρχουν δύο κλάσεις, $Σχετικό$ και $Μη\ Σχετικό$. Με άλλα λόγια, ένα σύστημα ανάκτησης πληροφορίας μπορεί να θεωρηθεί

ως ένας κατηγοριοποιητής δύο κλάσεων, το οποίο επιχειρεί να θέσει μία ετικέτα $Σχετικό$ ή $Μη\ Σχετικό$ σε κάθε έγγραφο σε ένα σύνολο εγγράφων. Δηλαδή, κάνει ανάκτηση το υποσύνολο των εγγράφων, τα οποία πιστεύει ότι είναι σχετικά.

Η μετρική που χρησιμοποιήθηκε στα πλαίσια των πειραμάτων της εργασίας είναι η ακρίβεια αποτελεσμάτων, η οποία προσαρμόστηκε έτσι ώστε να αντιστοιχεί στην ιδιαιτερότητα της εφαρμογής της εργασίας. Συγκεκριμένα, η ακρίβεια στα πλαίσια της εργασίας ορίζεται ως

$$Ακρίβεια = \frac{Ολοκληρώθηκε Επιτυχώς}{Ολοκληρώθηκε Επιτυχώς + Ολοκληρώθηκε Μη Επιτυχώς}$$

Η μετρική $Ολοκληρώθηκε Επιτυχώς$ καθορίζει τον αριθμό των χρονικών κανόνων συσχέτισης του προγράμματος, οι οποίοι επαληθεύθηκαν επιτυχώς κατά τη διαδικασία ελέγχου. Συγκεκριμένα, είναι ο αριθμός των κανόνων στους οποίους επαληθεύθηκε τόσο το σώμα του κανόνα, όσο και η κεφαλή του κανόνα. Η μετρική $Ολοκληρώθηκε Μη Επιτυχώς$ καθορίζει τον αριθμό των χρονικών κανόνων συσχέτισης του προγράμματος, για τους οποίους δεν υπήρχε επιτυχής επαλήθευση κατά τη διαδικασία ελέγχου. Συγκεκριμένα, είναι ο αριθμός των κανόνων, στους οποίους επαληθεύθηκε το σώμα του κανόνα, αλλά για κάποιο στοιχείο της κεφαλής του κανόνα δεν υπήρχε επιτυχής επαλήθευση.

7.4 Περιγραφή πειραμάτων

Τα πειράματα που εκτελέστηκαν στα πλαίσια της εργασίας βασίζονται σε θεωρία και αλγορίθμους, οι οποίοι περιεγράφηκαν στις παραπάνω ενότητες. Από τη θεωρία που περιγράφει την ανίχνευση συμβάντων σε ροές δεδομένων αισθητήρων πολλαπλών μεταβλητών, χρησιμοποιήθηκαν ο αλγόριθμος συσσωρευτικού αθροίσματος και ο αλγόριθμος διαγραμμάτων ελέγχου Shewhart. Το συστατικό της υλοποίησης που υλοποιεί την ανίχνευση συμβάντων σε ροές δεδομένων αισθητήρων πολλαπλών μεταβλητών δέχεται ως είσοδο ένα διάνυσμα πλαισίου και δίνει ως έξοδο ένα διάνυσμα συμβάντων, σε κάθε χρονικό βήμα. Στη συνέχεια, από τη θεωρία που περιγράφει τη συσχέτιση συμβάντων σε ροές δεδομένων συμβάντων πολλαπλών μεταβλητών, χρησιμοποιήθηκε ο αλγόριθμος μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών, ο οποίος δέχεται ως είσοδο, σε κάθε χρονικό βήμα, ένα διάνυσμα συμβάντων που προέκυψε από την ανίχνευση συμβάντων σε ροές δεδομένων αισθητήρων πολλαπλών μεταβλητών. Τέλος, από τη θεωρία που περιγράφει το

προσαρμοστικό φιλτράρισμα εξαρτήσεων συμβάντων, χρησιμοποιήθηκαν οι συναρτήσεις γραμμικής και εκθετικής απόσβεσης. Οι προς διερεύνηση κανόνες είναι χρονικοί κανόνες συσχέτισης, και συγκεκριμένα διασυναλλαγικοί χρονικοί κανόνες συσχέτισης, όπου η δομή του κανόνα μπορεί να έχει κάποια από τις ακόλουθες δομές κανόνων συσχέτισης:

Πίνακας 4: Πιθανές δομές χρονικών κανόνων συσχέτισης, οι οποίοι χρησιμοποιήθηκαν στα πλαίσια της εργασίας

Δομή κανόνα	**Παράδειγμα**
$μεμονωμένο \rightarrow μεμονωμένο$	$X \rightarrow \Psi$
$μεμονωμένο \rightarrow πολλαπλά$	$X \rightarrow \Psi, Z$
$πολλαπλά \rightarrow μεμονωμένο$	$X, \Psi \rightarrow Z$
$πολλαπλά \rightarrow πολλαπλά$	$X, \Psi \rightarrow Z, Y$

Ο αριθμός των στοιχείων στο σώμα του κανόνα καθορίζεται από τη μετρική m του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών, ενώ ο αριθμός των στοιχείων στην κεφαλή του κανόνα καθορίζεται από τη μετρική l του αλγορίθμου.

Οι παράμετροι του κάθε πειράματος καθορίζονται στην αρχή του πειράματος μέσω ενός αρχείου ιδιοτήτων. Στη συνέχεια, γίνεται επεξήγηση των παραμέτρων αυτών, και παρουσιάζονται οι πίνακες, γραφικές παραστάσεις, παρατηρήσεις και συμπεράσματα για τα πειράματα που εκτελέστηκαν στα πλαίσια της εργασίας.

- **NumOfIterations**

Ο αριθμός των χρονικών βημάτων που λαμβάνονται υπόψη από το πρόγραμμα, δηλαδή ο αριθμός των γραμμών από το αρχείο εισόδου που θα δοθούν ως είσοδος στο πρόγραμμα.

- **m**

Η παράμετρος m του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών.

- **l**

Η παράμετρος l του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών.

- **EventsIterationThreshold**

Μία τιμή κατωφλίου, η οποία χρησιμοποιήθηκε για εξοικονόμηση πόρων του συστήματος, και αύξηση ταχύτητας εκτέλεσης. Εάν σε ένα χρονικό βήμα εμφανιστεί αριθμός συμβάντων που ξεπερνάει το κατώφλι αυτό, αγνοείται το χρονικό βήμα. Μία τέτοια ενέργεια βασίζεται στην ευριστική ότι ένας μεγάλος αριθμός από συμβάντα σε ένα χρονικό βήμα καθορίζει σίγουρα κάποιο γενικό συναγερμό, ο οποίος όμως επηρεάζει μεγάλο αριθμό από μετρήσεις, και η χρησιμότητα του στα πλαίσια της διερεύνησης είναι αμελητέα.

- **kFixedCombinations**

Μία τιμή κατωφλίου, η οποία καθορίζει το μέγιστο αριθμό από συμβάντα που λαμβάνονται υπόψη για κάθε στοιχείο του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών, δηλαδή ο μέγιστος αριθμός συμβάντων που μπορούν είναι εμφανή στο ίδιο κόμβο κάποιου δένδρου στην ιστορική δομή του αλγορίθμου.

- **StartTestingIteration**

Το χρονικό βήμα, στο οποίο εκκινεί η διαδικασία ελέγχου. Η τιμή αυτή πρέπει να είναι τέτοια, ώστε όταν εκκινεί η διαδικασία ελέγχου, το μοντέλο εκπαίδευσης του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών να είναι σε ένα ικανοποιητικό βαθμό ωρίμανσης.

- **EventDetectionAlgorithm**

Ο αλγόριθμος, ο οποίος χρησιμοποιείται για την ανίχνευση συμβάντων σε ροές δεδομένων αισθητήρων πολλαπλών μεταβλητών. Οι πιθανοί αλγόριθμοι είναι ο αλγόριθμος συσσωρευτικού αθροίσματος ή ο αλγόριθμος διαγραμμάτων ελέγχου Shewhart.

- **AgingFunction**

Η συνάρτηση που χρησιμοποιήθηκε στα πλαίσια της διαδικασίας προσαρμοστικού φιλτραρίσματος εξαρτήσεων συμβάντων. Υπάρχει η επιλογή μη εφαρμογής κάποιας διαδικασίας προσαρμοστικού φιλτραρίσματος εξαρτήσεων συμβάντων. Στην περίπτωση εφαρμογής προσαρμοστικού φιλτραρίσματος εξαρτήσεων συμβάντων, οι πιθανές προσεγγίσεις που μπορούν να χρησιμοποιηθούν είναι η συνάρτηση γραμμικής απόσβεσης και η συνάρτηση εκθετικής απόσβεσης.

- **kAgingValue**

Η τιμή k της συνάρτησης απόσβεσης που χρησιμοποιήθηκε στα πλαίσια της διαδικασίας προσαρμοστικού φιλτραρίσματος εξαρτήσεων συμβάντων. Μία τυπική τιμή της μεταβλητής k είναι $0{,}1$, και όσο αυξάνεται η τιμή της μεταβλητής, τόσο πιο απότομα γίνεται απόσβεση της συνάρτησης. Η μεταβλητή k είναι μεταβλητή τόσο της γραμμικής συνάρτησης απόσβεσης, όσο και της εκθετικής συνάρτησης απόσβεσης.

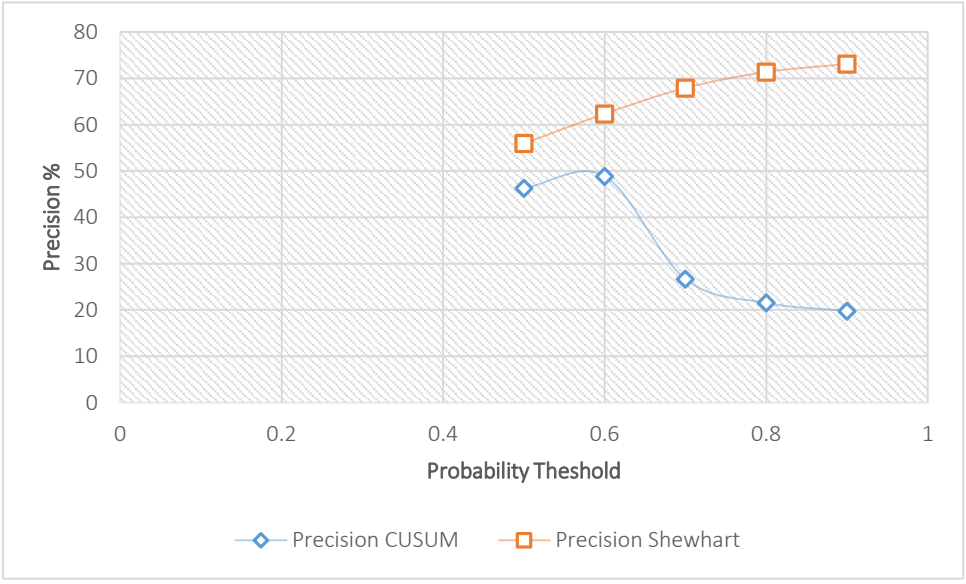
- **ProbabilityThreshold**

Μία τιμή κατωφλίου πιθανότητας, η οποία παίζει καθοριστικό ρόλο στην αποτελεσματικότητα του αλγορίθμου μεταβλητής τάξης συσχέτισης συμβάντων πολλαπλών μεταβλητών. Εάν η πιθανότητα ενός χρονικού κανόνα συσχέτισης, ο οποίος προέκυψε από την αποδόμηση κάποιου δένδρου στην ιστορική δομή του αλγορίθμου ξεπερνάει την τιμή κατωφλίου, τότε ο κανόνας αυτός εισάγεται σε ένα στιγμιότυπο της κλάσης $ΣυνδυασμόςΕλέγχου$. Με άλλα λόγια, εάν η τιμή πιθανότητας ενός κανόνα ξεπερνάει την τιμή κατωφλίου, τότε ο κανόνας αυτός εισάγεται στη διαδικασία ελέγχου και δυνητικά συμμετέχει στο προσδιορισμό της τιμής ακρίβειας του πειράματος.

7.4.1 Πειράματα μεταβολής τιμής κατωφλίου πιθανότητας

Τα πειράματα της υποενότητας αυτής εκτελέστηκαν με σταθερές τιμές όλων των παραμέτρων, εκτός από την παράμετρο του κατωφλίου πιθανότητας. Η μελέτη των πειραμάτων ομαδοποιείται για κοινές τιμές παραμέτρων m και l του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων πολλαπλών μεταβλητών, και με διαφοροποίηση του αλγορίθμου ανίχνευσης συμβάντων σε ροές δεδομένων αισθητήρων πολλαπλών μεταβλητών.

Είσοδος: $numOfIterations = 8000$, $m = 1$, $l = 1$, $eventsIterationThreshold = 5$, $kFixedCombinations = 3$, $startTestingIteration = 100$, $agingFunction = NONE$.



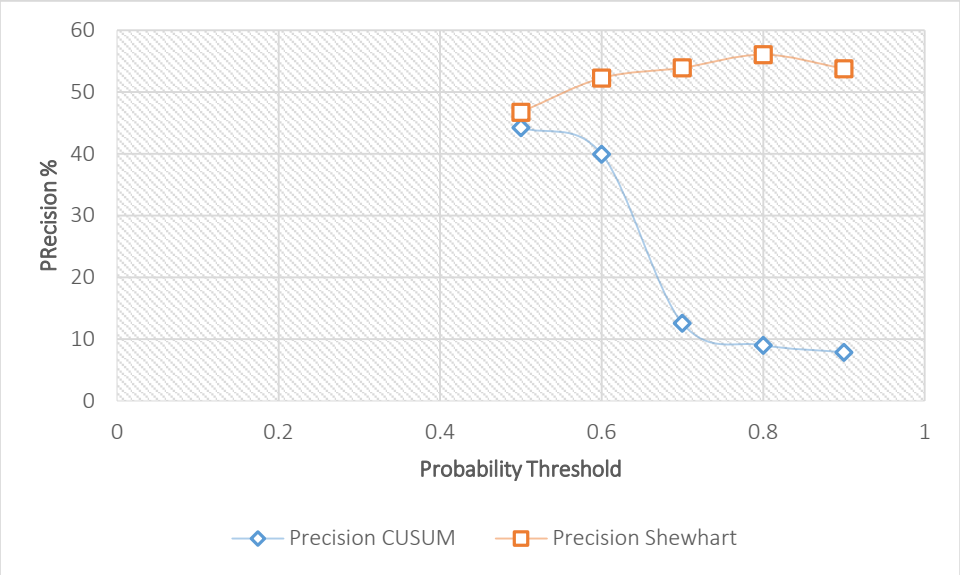
PROBABILITY THRESHOLD	PRECISION CUSUM	PRECISION SHEWHART
0.5	46.30505624	55.93504778
0.6	48.86459209	62.33390967
0.7	26.67249672	67.93237251
0.8	21.5230037	71.37115964
0.9	19.77272727	73.10104033

Γραφική παράσταση 1: Πειράματα μεταβολής τιμής κατωφλίου πιθανότητας για m=1, l=1

Οι τιμές ακρίβειας του αλγορίθμου συσσωρευτικού αθροίσματος παρουσιάζουν πτώση, όσο αυξάνεται η τιμή κατωφλίου πιθανότητας, συμπεριφορά η οποία είναι αντίστροφη του αναμενόμενου. Αυτό μπορεί να ερμηνευθεί ως μη καταλληλότητα του αλγορίθμου συσσωρευτικού αθροίσματος για ανίχνευσης συμβάντων από τα δεδομένα αισθητήρων στην είσοδο, λόγω του ότι ο αλγόριθμος λαμβάνει υπόψη την υπόθεση ότι τα δεδομένα εισόδου ακολουθούν κανονική κατανομή, γεγονός που δεν είναι αληθές για τα δεδομένα εισόδου που χρησιμοποιήθηκαν. Ωστόσο, επίσης αξίζει να σημειωθεί ότι ο αλγόριθμος συσσωρευτικού αθροίσματος έχει αρκετά καλές τιμές ακρίβειας για τιμές $0,5$ και $0,6$ του κατωφλίου πιθανότητας, σε σύγκριση με μεγαλύτερες τιμές κατωφλίου.

Οι τιμές ακρίβειας του αλγορίθμου διαγραμμάτων ελέγχου Shewhart παρουσιάζουν μία προσεγγιστικά γραμμική αύξηση όσο αυξάνεται η τιμή κατωφλίου πιθανότητας, συμπεριφορά η οποία είναι αναμενόμενη. Επίσης αξίζει να σημειωθεί ότι πατατηρούνται ικανοποιητικές τιμές ακρίβειας, που για τιμή κατωφλίου πιθανότητας $0,9$ είναι περίπου 73%. Αυτό μπορεί να ερμηνευθεί ως καταλληλότητα του αλγορίθμου διαγραμμάτων ελέγχου Shewhart για ανίχνευση συμβάντων από τα δεδομένα αισθητήρων στην είσοδο, λόγω του ότι ο αλγόριθμος δεν λαμβάνει υπόψη κάποια υπόθεση για την κατανομή των δεδομένων εισόδου. Τέλος, από τις υψηλές τιμές των τιμών ακρίβειας, φαίνεται η αποτελεσματικότητα του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών. Συγκεκριμένα, λόγω του αριθμού των m και l, τα οποία δεν είναι πολύ μεγάλα, τα δένδρα στην ιστορική δομή δεν είναι μεγάλης έκτασης. Συνεπώς, το μοντέλο εκπαίδευσης του αλγορίθμου είναι αρκετά γενικό, και έτσι η αποτελεσματικότητα του αλγορίθμου μπορεί να χαρακτηριστεί ως πολύ καλή.

Είσοδος: $numOfIterations = 8000$, $m = 1$, $l = 2$, $eventsIterationThreshold = 5$, $kFixedCombinations = 3$, $startTestingIteration = 100$, $agingFunction = NONE$.



PROBABILITY THRESHOLD	PRECISION CUSUM	PRECISION SHEWHART
0.5	44.25655224	46.75535499
0.6	39.95196451	52.26945818
0.7	12.56313871	53.92807514
0.8	9.035392063	56.05524617
0.9	7.872044208	53.80344463

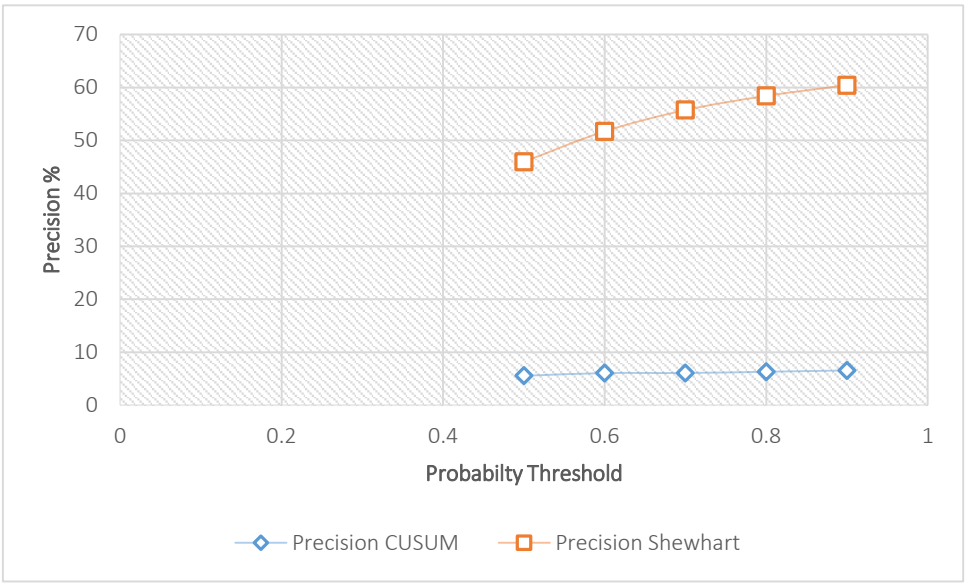
Γραφική παράσταση 2: Πειράματα μεταβολής τιμής κατωφλίου πιθανότητας για m=1, l=2

Οι τιμές ακρίβειας του αλγορίθμου συσσωρευτικού αθροίσματος παρουσιάζουν πτώση όσο αυξάνεται η τιμή κατωφλίου πιθανότητας, συμπεριφορά η οποία είναι αντίστροφη του αναμενόμενου. Αυτό, όπως προηγουμένως, μπορεί να ερμηνευθεί ως μη καταλληλότητα του αλγορίθμου για ανίχνευση συμβάντων από τα δεδομένα αισθητήρων στην είσοδο, λόγω του ότι ο αλγόριθμος λαμβάνει υπόψη την υπόθεση ότι τα δεδομένα εισόδου ακολουθούν κανονική κατανομή. Οι τιμές ακρίβειας και στους δύο αλγόριθμους παρουσιάζουν μία μικρή πτώση σε σύγκριση με τιμές $m = 1$ και $l = 1$, γεγονός που μπορεί να ερμηνευτεί ως κάποια συγκεκριμενοποίηση του μοντέλου εκπαίδευσης του

αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών, συγκεκριμενοποίηση η οποία επηρεάζει αρνητικά την αποτελεσματικότητα του αλγορίθμου.

Οι τιμές ακρίβειας του αλγορίθμου διαγραμμάτων ελέγχου Shewhart παρουσιάζουν μία προσεγγιστικά γραμμική αύξηση όσο αυξάνεται η τιμή κατωφλίου πιθανότητας, συμπεριφορά η οποία είναι αναμενόμενη. Αυτό μπορεί, όπως προηγουμένως, να ερμηνευθεί ως καταλληλότητα του αλγορίθμου για ανίχνευση συμβάντων από τα δεδομένα αισθητήρων στην είσοδο, λόγω του ότι ο αλγόριθμος δεν λαμβάνει υπόψη κάποια υπόθεση για την κατανομή των δεδομένων εισόδου. Τέλος, από τις υψηλές τιμές των τιμών ακρίβειας, φαίνεται η αποτελεσματικότητα του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών.

Είσοδος: $numOfIterations = 8000$, $m = 2$, $l = 1$, $eventsIterationThreshold = 5$, $kFixedCombinations = 3$, $startTestingIteration = 100$, $agingFunction = NONE$.



PROBABILITY THRESHOLD	PRECISION CUSUM	PRECISION SHEWHART
0.5	5.550085493	45.97212215
0.6	6.06295758	51.72185152
0.7	6.083668622	55.74431184
0.8	6.311980702	58.42019307

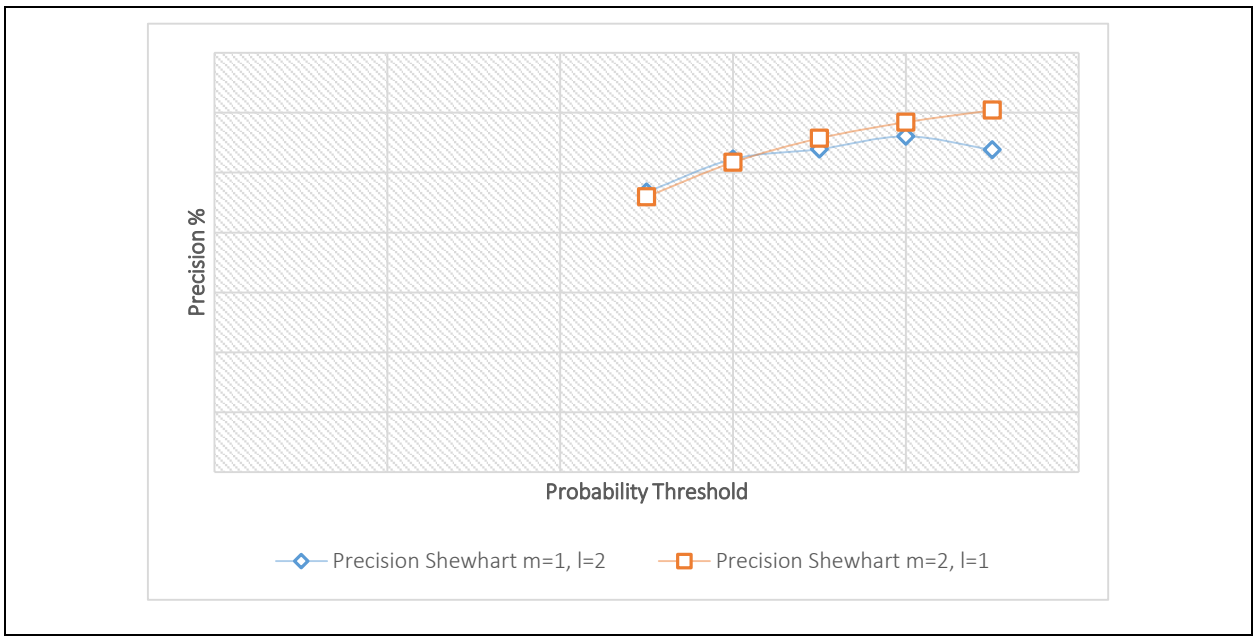
	0.9	6.571441783	60.43106873

Γραφική παράσταση 3: Πειράματα μεταβολής τιμής κατωφλίου πιθανότητας για m=2, l=1

Οι τιμές ακρίβειας του αλγορίθμου συσσωρευτικού αθροίσματος παρουσιάζουν μία προσεγγιστικά σταθερή συμπεριφορά όσο αυξάνεται η τιμή κατωφλίου πιθανότητας, συμπεριφορά η οποία είναι μη αναμενόμενη. Επίσης, οι τιμές ακρίβειας είναι για όλες τις τιμές κατωφλίου πιθανότητας χαμηλές. Αυτό, όπως προηγουμένως, μπορεί να ερμηνευθεί ως μη καταλληλότητα του αλγορίθμου για ανίχνευση συμβάντων από τα δεδομένα αισθητήρων στην είσοδο, λόγω του ότι ο αλγόριθμος λαμβάνει υπόψη την υπόθεση ότι τα δεδομένα εισόδου ακολουθούν κανονική κατανομή. Οι τιμές ακρίβειας και στους δύο αλγόριθμους παρουσιάζουν μία μικρή πτώση σε σύγκριση με τιμές $m = 1$ και $l = 1$, γεγονός που μπορεί να ερμηνευτεί ως κάποια συγκεκριμενοποίηση του μοντέλου εκπαίδευσης του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών, συγκεκριμενοποίηση η οποία επηρεάζει αρνητικά την αποτελεσματικότητα του αλγορίθμου.

Οι τιμές ακρίβειας του αλγορίθμου διαγραμμάτων ελέγχου Shewhart παρουσιάζουν μία προσεγγιστικά γραμμική αύξηση όσο αυξάνεται η τιμή κατωφλίου πιθανότητας, συμπεριφορά η οποία είναι αναμενόμενη. Αυτό μπορεί, όπως προηγουμένως, να ερμηνευθεί ως καταλληλότητα του αλγορίθμου για ανίχνευση συμβάντων από τα δεδομένα αισθητήρων στην είσοδο, λόγω του ότι ο αλγόριθμος δεν λαμβάνει υπόψη κάποια υπόθεση για την κατανομή των δεδομένων εισόδου. Τέλος, από τις υψηλές τιμές των τιμών ακρίβειας στην περίπτωση ανίχνευσης συμβάντων με τον αλγόριθμο διαγραμμάτων ελέγχου Shewhart, φαίνεται η αποτελεσματικότητα του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών.

Είσοδος: $numOfIterations = 8000$, $eventsIterationThreshold = 5$, $kFixedCombinations = 3$, $startTestingIteration = 100$, $eventDetectionAlgorithm = SHEWHART$, $agingFunction = NONE$.



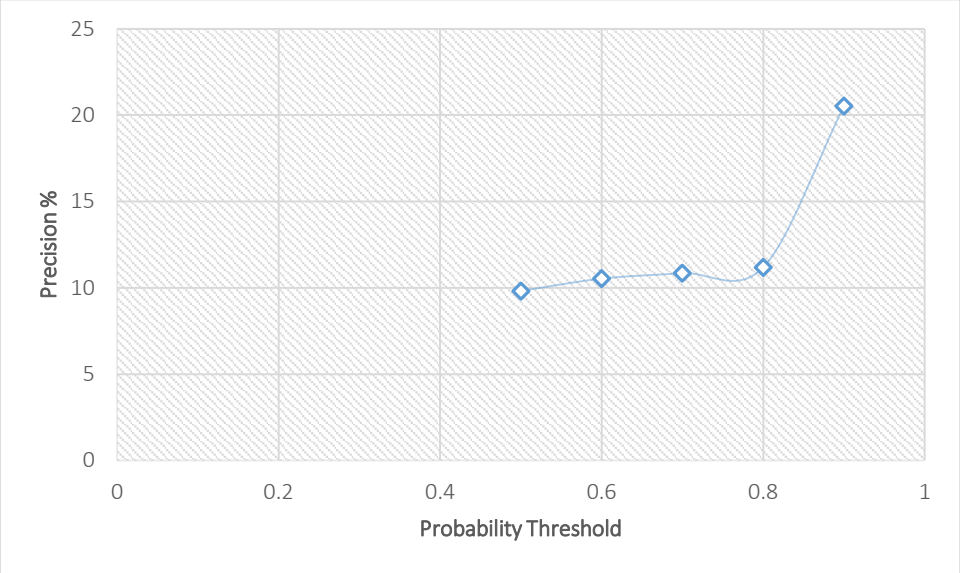
Γραφική παράσταση 4: Πειράματα μεταβολής τιμής κατωφλίου πιθανότητας με ανίχνευση συμβάντων με τον αλγόριθμο διαγραμμάτων ελέγχου Shewhart

Συγκρίνοντας τη συμπεριφορά με ανίχνευση συμβάντων με τον αλγόριθμο διαγραμμάτων ελέγχου Shewhart για τις ίδιες παραμέτρους και ίδιο ύψος δένδρου, οι τιμές ακρίβειας έχουν ικανοποιητικά καλές τιμές. Στην περίπτωση που το μοντέλο του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών έχει τιμές παραμέτρων $m = 2$ και $l = 1$, η τιμή ακρίβειας είναι ελάχιστα πιο μεγάλη σε σύγκριση με το μοντέλο που έχει τιμές παραμέτρων $m = 1$ και $l = 2$, ενώ όσο αυξάνεται η τιμή κατωφλίου πιθανότητας, ο αλγόριθμος στην πρώτη περίπτωση φαίνεται να έχει καλύτερη αποτελεσματικότητα. Αυτό ερμηνεύεται μέσω της μεγαλύτερης τιμής της παραμέτρου m. Όσο μεγαλώνει η τιμή της παραμέτρου m, τόσο μειώνεται ο αριθμός των χρονικών κανόνων συσχέτισης που υπόκεινται σε έλεγχο κατά τη διαδικασία ελέγχου, αφού στην περίπτωση που δεν επαληθεύεται ολόκληρο το σώμα του κανόνα, ο κανόνας δεν καταμετρείται για τον υπολογισμό της ακρίβειας.

Να σημειωθεί ότι στην περίπτωση $l > 1$, οι κανόνες που υπόκεινται σε έλεγχο κατά τη διαδικασία ελέγχου έχουν αριθμό στοιχείων στην κεφαλή που ανήκει στο διάστημα $[1, l]$. Με άλλα λόγια, εάν ισχύει $l = 2$, τότε οι κανόνες που υπόκεινται σε έλεγχο κατά τη διαδικασία ελέγχου έχουν αριθμό στοιχείων στην κεφαλή, ο οποίος ανήκει στο διάστημα $[1,2]$. Αυτό μπορεί να εξηγήσει τις μεγάλες τιμές ακρίβειας του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών, αφού με αυτό τον τρόπο αυξάνεται περισσότερο ο αριθμός των χρονικών κανόνων συσχέτισης, οι οποίοι

χαρακτηρίζονται ως $Ολοκληρώθηκε Επιτυχώς$, από τον αριθμό των χρονικών κανόνων συσχέτισης, οι οποίοι χαρακτηρίζονται ως $Ολοκληρώθηκε Μη Επιτυχώς$.

Είσοδος: $numOfIterations = 7500$, $m = 2$, $l = 2$, $eventsIterationThreshold = 3$, $kFixedCombinations = 0$, $startTestingIteration = 100$, $agingFunction = NONE$.



PROBABILITY THRESHOLD	PRECISION SHEWHART
0.5	9.8213059
0.6	10.53284016
0.7	10.84777882
0.8	11.18745516
0.9	20.52142474

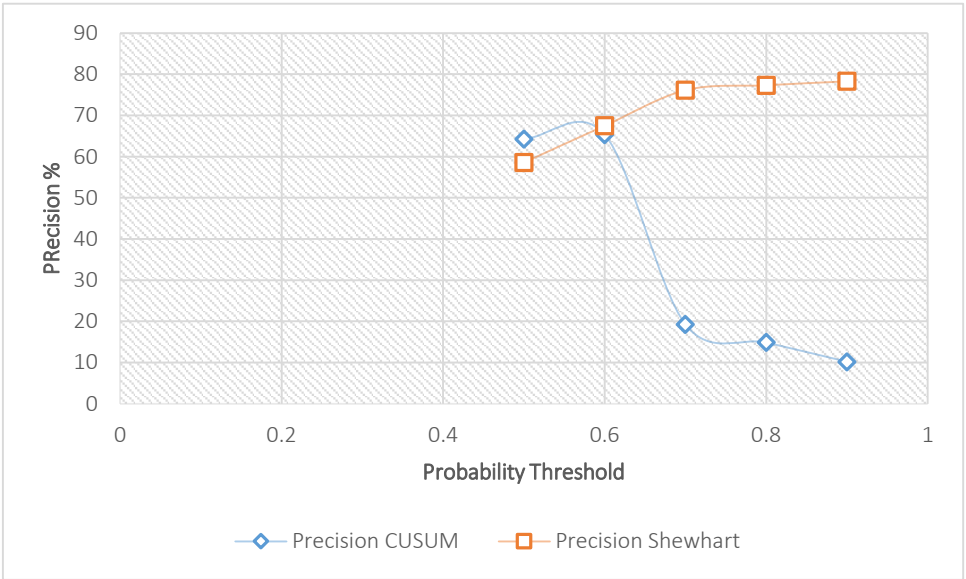
Γραφική παράσταση 5: Πειράματα μεταβολής τιμής κατωφλίου πιθανότητας για m=2, l=2

Να σημειωθεί ότι λόγω της αρκετά χαμηλής τιμής ακρίβειας στην περίπτωση ανίχνευσης συμβάντων με τον αλγόριθμο συσσωρευτικού αθροίσματος, για τιμές παραμέτρων του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών $m = 2$ και $l = 1$, θεωρήθηκε μη χρήσιμο η περαιτέρω μελέτη πειραμάτων για τιμές $l > 1$ με ανίχνευση συμβάντων με τον αλγόριθμο συσσωρευτικού αθροίσματος. Αυτό οφείλεται στο γεγονός ότι για τιμές $l > 1$ υπάρχει συγκεκριμενοποίηση του

αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών, συγκεκριμενοποίηση η οποία οδηγεί σε περαιτέρω μείωση της τιμής ακρίβειας.

Σε αυτή την ομάδα πειραμάτων παρατηρείται μία σχετικά χαμηλή τιμή ακρίβειας, η οποία αυξάνεται προσεγγιστικά γραμμικά μέχρι τιμή κατωφλίου πιθανότητας $0{,}8$, ενώ στη συνέχεια αυξάνεται με ραγδαίο ρυθμό. Η συμπεριφορά αυτή είναι αναμενόμενη, μιας και η αύξηση της τιμής κατωφλίου πιθανότητας περιορίζει τον αριθμό των κανόνων προς έλεγχο, κάνοντας επιλογή των κανόνων, οι οποίοι έχουν μεγαλύτερη πιθανότητα. Οι τιμές ακρίβειας παρουσιάζουν μία μικρή πτώση σε σύγκριση με τιμές προηγούμενων πειραμάτων, γεγονός που μπορεί να ερμηνευτεί ως κάποια συγκεκριμενοποίηση του μοντέλου εκπαίδευσης του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών, συγκεκριμενοποίηση η οποία επηρεάζει αρνητικά την αποτελεσματικότητα του αλγορίθμου.

Είσοδος: $numOfIterations = 7500$, $m = 1$, $l = 3$, $eventsIterationThreshold = 3$, $kFixedCombinations = 0$, $startTestingIteration = 100$, $agingFunction = NONE$.



PROBABILITY THRESHOLD	PRECISION CUSUM	PRECISION SHEWHART
0.5	64.28573585	58.56000153
0.6	65.34009495	67.48206408
0.7	19.26461971	76.12281363

0.8	14.8600694	77.33270702
0.9	10.18869874	78.2851188

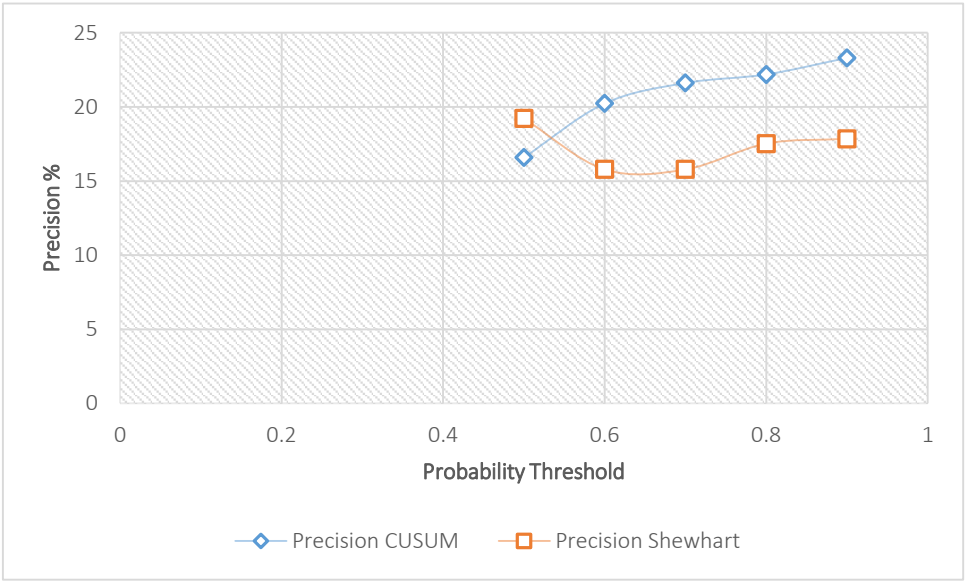
Γραφική παράσταση 6: Πειράματα μεταβολής τιμής κατωφλίου πιθανότητας για m=1, l=3

Οι τιμές ακρίβειας του αλγορίθμου συσσωρευτικού αθροίσματος παρουσιάζουν πτώση όσο αυξάνεται η τιμή κατωφλίου πιθανότητας, συμπεριφορά η οποία είναι αντίστροφη του αναμενόμενου. Αυτό, όπως προηγουμένως, μπορεί να ερμηνευθεί ως μη καταλληλότητα του αλγορίθμου συσσωρευτικού αθροίσματος για ανίχνευση συμβάντων από τα δεδομένα αισθητήρων στην είσοδο, λόγω του ότι ο αλγόριθμος λαμβάνει υπόψη την υπόθεση ότι τα δεδομένα εισόδου ακολουθούν κανονική κατανομή. Ωστόσο, αξίζει επίσης να σημειωθεί ότι ο αλγόριθμος συσσωρευτικού αθροίσματος έχει αρκετά καλές τιμές ακρίβειας για τιμές $0,5$ και $0,6$ του κατωφλίου πιθανότητας, σε σύγκριση με μεγαλύτερες τιμές κατωφλίου. Οι τιμές ακρίβειας και στους δύο αλγόριθμους παρουσιάζουν μία μικρή πτώση σε σύγκριση με τιμές $l < 3$, γεγονός που μπορεί να ερμηνευτεί ως κάποια συγκεκριμενοποίηση του μοντέλου εκπαίδευσης του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών, συγκεκριμενοποίηση η οποία επηρεάζει αρνητικά την αποτελεσματικότητα του αλγορίθμου.

Οι τιμές ακρίβειας του αλγορίθμου διαγραμμάτων ελέγχου Shewhart παρουσιάζουν μία προσεγγιστικά γραμμική αύξηση όσο αυξάνεται η τιμή κατωφλίου πιθανότητας, συμπεριφορά η οποία είναι αναμενόμενη. Αυτό μπορεί, όπως προηγουμένως, να ερμηνευθεί ως καταλληλότητα του αλγορίθμου για ανίχνευση συμβάντων από τα δεδομένα αισθητήρων στην είσοδο, λόγω του ότι ο αλγόριθμος δεν λαμβάνει υπόψη κάποια υπόθεση για την κατανομή των δεδομένων εισόδου. Τέλος, από τις υψηλές τιμές των τιμών ακρίβειας, φαίνεται η αποτελεσματικότητα του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών. Οι τιμές ακρίβειας του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών με ανίχνευση συμβάντων με τον αλγόριθμο διαγραμμάτων ελέγχου Shewhart είναι οι μέγιστες σε σύγκριση με όλα τα προηγούμενα πειράματα. Αυτό ερμηνεύεται ως η κατάλληλη γενίκευση του αλγόριθμου, η οποία επιτυγχάνεται με την ποικιλομορφία των χρονικών κανόνων συσχέτισης, ποικιλομορφία που προκύπτει από το μεταβλητό μήκος στοιχείων στην κεφαλή των χρονικών κανόνων συσχέτισης. Αυτό οφείλεται, όπως αναφέρθηκε προηγουμένως, στο γεγονός ότι για τιμές παραμέτρου $l >$

1, οι χρονικοί κανόνες συσχέτισης μπορούν να έχουν μεταβλητό αριθμό στοιχείων στην κεφαλή που ανήκει στο διάστημα $[1, l]$.

Είσοδος: $numOfIterations = 3500\ CUSUM$, $numOfIterations = 7500\ Shewhart$, $m = 3$, $l = 1$, $eventsIterationThreshold = 3$, $kFixedCombinations = 0$, $startTestingIteration = 100$, $agingFunction = NONE$.



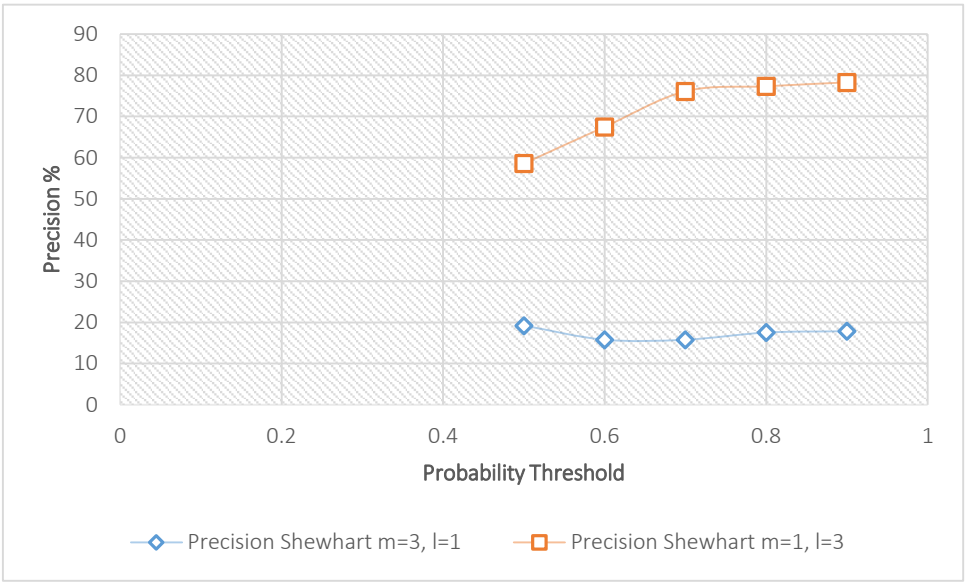
PROBABILITY THRESHOLD	PRECISION CUSUM	PRECISION SHEWHART
0.5	16.58587499	19.21504784
0.6	20.23312756	15.78343313
0.7	21.61493974	15.78091311
0.8	22.19136819	17.52706451
0.9	23.30431261	17.85071493

Γραφική παράσταση 7: Πειράματα μεταβολής τιμής κατωφλίου πιθανότητας για m=3, l=1

Οι τιμές ακρίβειας του αλγορίθμου συσσωρευτικού αθροίσματος παρουσιάζουν αύξηση όσο αυξάνεται η τιμή κατωφλίου πιθανότητας, συμπεριφορά η οποία είναι αναμενόμενη. Οι τιμές ακρίβειας του αλγορίθμου διαγραμμάτων ελέγχου Shewhart παρουσιάζουν μία κυμαινόμενη συμπεριφορά, όσο αυξάνεται η τιμή κατωφλίου πιθανότητας. Οι τιμές ακρίβειας και στους δύο αλγόριθμους παρουσιάζουν πτώση σε σύγκριση με τιμές $m < 3$,

γεγονός που μπορεί να ερμηνευτεί ως κάποια συγκεκριμενοποίηση του μοντέλου εκπαίδευσης του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών, συγκεκριμενοποίηση η οποία επηρεάζει αρνητικά την αποτελεσματικότητα του αλγορίθμου.

Είσοδος: $numOfIterations = 7500$, $eventsIterationThreshold = 3$, $kFixedCombinations = 0$, $startTestingIteration = 100$, $eventDetectionAlgorithm = SHEWHART$, $agingFunction = NONE$.



Γραφική παράσταση 8: Πειράματα μεταβολής τιμής κατωφλίου πιθανότητας με ανίχνευση συμβάντων με τον αλγόριθμο διαγραμμάτων ελέγχου Shewhart

Συγκρίνοντας τη συμπεριφορά με ανίχνευση συμβάντων με τον αλγόριθμο διαγραμμάτων ελέγχου Shewhart με τις ίδιες παραμέτρους και ίδιο ύψος δένδρου, οι τιμές ακρίβειας διαφέρουν αρκετά. Στην περίπτωση που το μοντέλο του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών έχει τιμές παραμέτρων $m = 3$ και $l = 1$, η τιμή ακρίβειας είναι σχετικά χαμηλή, σε σύγκριση με τον αλγόριθμο που έχει τιμές παραμέτρων $m = 1$ και $l = 3$. Όσο αυξάνεται η τιμή κατωφλίου πιθανότητας στην πρώτη περίπτωση, ο αλγόριθμος φαίνεται να έχει σταθερή συμπεριφορά, ενώ στη δεύτερη περίπτωση αυξάνεται η αποτελεσματικότητα του αλγορίθμου.

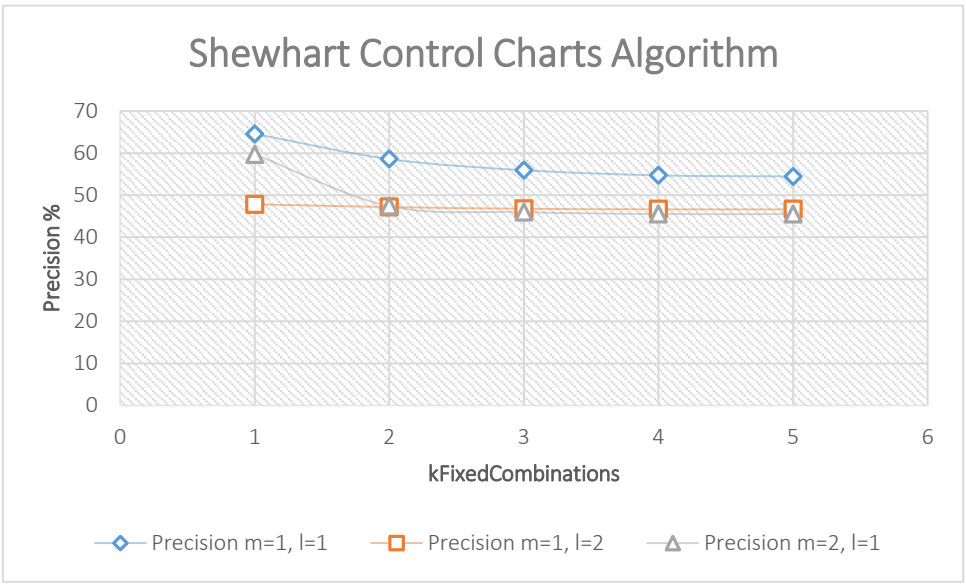
Στην πρώτη περίπτωση η συμπεριφορά ερμηνεύεται ότι οφείλεται στη μεγαλύτερη τιμή της παραμέτρου m. Όσο μεγαλώνει η τιμή της παραμέτρου m, τόσο μειώνεται ο αριθμός των χρονικών κανόνων συσχέτισης οι οποίοι υπόκεινται σε έλεγχο κατά τη διαδικασία ελέγχου, αφού στην περίπτωση που δεν επαληθεύεται ολόκληρο το σώμα του

κανόνα, ο κανόνας δεν καταμετρείται για τον υπολογισμό της ακρίβειας. Συνεπώς, υπάρχει συγκεκριμενοποίηση του μοντέλου εκπαίδευσης, η οποία οδηγεί σε μείωση της αποτελεσματικότητας του αλγορίθμου. Στη δεύτερη περίπτωση δεν υπάρχει συγκεκριμενοποίηση του μοντέλου, αφού όπως αναφέρθηκε προηγουμένως, για τιμές παραμέτρου στον αλγόριθμο μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών $l > 1$, οι κανόνες που υπόκεινται σε έλεγχο κατά τη διαδικασία ελέγχου έχουν αριθμό στοιχείων στην κεφαλή που ανήκει στο διάστημα $[1, l]$. Αυτό μπορεί να εξηγήσει τις μεγάλες τιμές ακρίβειας του αλγορίθμου, αφού με αυτό τον τρόπο αυξάνεται περισσότερο ο αριθμός των χρονικών κανόνων συσχέτισης, οι οποίοι χαρακτηρίζονται ως $Ολοκληρώθηκε Επιτυχώς$, από τον αριθμό των χρονικών κανόνων συσχέτισης, οι οποίοι χαρακτηρίζονται ως $Ολοκληρώθηκε Μη Επιτυχώς$.

7.4.2 Πειράματα μεταβολής τιμής σταθερού αριθμού κ συνδυασμών

Τα πειράματα της υποενότητας αυτής εκτελέστηκαν με σταθερές τιμές όλων των παραμέτρων, εκτός από την παράμετρο του σταθερού αριθμού κ συνδυασμών. Η μελέτη των πειραμάτων ομαδοποιείται για κοινές τιμές των παραμέτρων του συστήματος, και με διαφοροποίηση του αλγορίθμου ανίχνευσης συμβάντων σε ροές δεδομένων αισθητήρων πολλαπλών μεταβλητών.

Είσοδος: $numOfIterations = 8000$, $eventsIterationThreshold = 5$, $startTestingIteration = 100$, $eventDetectionAlgorithm = SHEWHART$, $agingFunction = NONE$, $probabilityThreshold = 0.5$.



SHEWHART CONTROL CHARTS ALGORITHM

KFIXEDCOMBINATIONS	Precision m=1, l=1	Precision m=1, l=2	Precision m=2, l=1
5	54.44329021	46.621183	45.49400134
4	54.70826825	46.62858836	45.516148
3	55.93504778	46.75535499	45.97212215
2	58.57255351	47.17433682	47.37808812
1	64.56644559	47.82222652	59.67564989

Γραφική παράσταση 9: Πειράματα μεταβολής τιμής σταθερού αριθμού κ συνδυασμών με ανίχνευση συμβάντων με τον αλγόριθμο διαγραμμάτων ελέγχου Shewhart

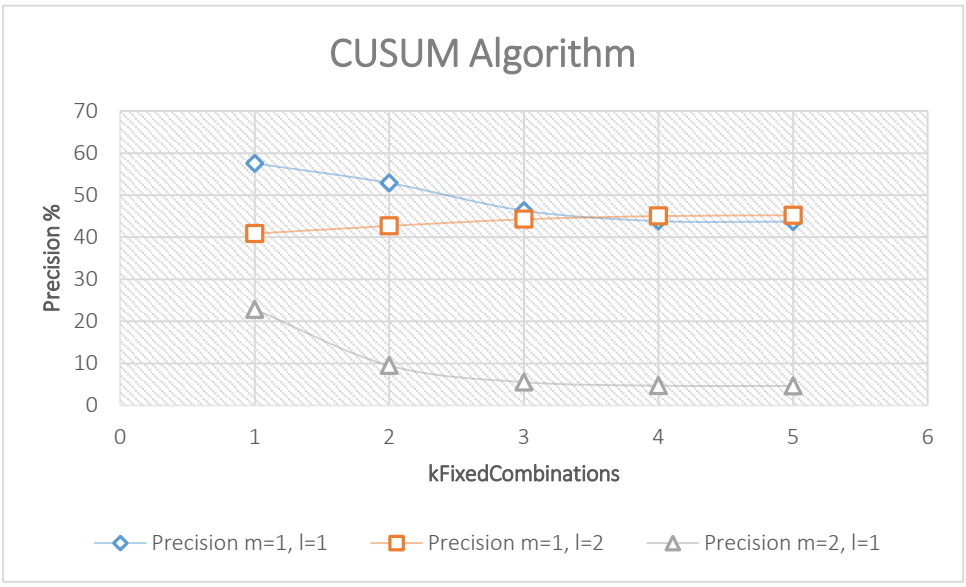
Τα πειράματα εκτελέστηκαν για τον ίδιο αλγόριθμο ανίχνευσης συμβάντων, τον αλγόριθμο διαγραμμάτων ελέγχου Shewhart και έγινε ομαδοποίηση των πειραμάτων για ίδιες τιμές m και l του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών, μεταβάλλοντας την τιμή του σταθερού αριθμού κ συνδυασμών. Οι τιμές του αριθμού κ συνδυασμών στα πειράματα ανήκουν στο διάστημα $[1,5]$. Από τα πειράματα, παρατηρείται και στις τρεις περιπτώσεις μία αύξηση της τιμής ακρίβειας, σε μεγάλο ή σε μικρό βαθμό, όσο μειώνεται ο αριθμός των κ συνδυασμών. Αυτό μπορεί να ερμηνευθεί ως κάποια γενίκευση του μοντέλου εκπαίδευσης του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών, γενίκευση η οποία αυξάνει την αποτελεσματικότητα του αλγορίθμου. Να σημειωθεί, όπως αναφέρθηκε και

προηγουμένως, ότι ο αριθμός κ συνδυασμών καθορίζει το μέγιστο αριθμό συνδυασμών, οι οποίοι εμφανίζονται σε ένα κόμβο κάποιου δένδρου της ιστορικής δομής του αλγορίθμου. Συνεπώς, με αύξηση του αριθμού κ συνδυασμών, το μοντέλο εκπαίδευσης του αλγορίθμου εξειδικεύεται, με αναμενόμενο αποτέλεσμα τη μείωση της αποτελεσματικότητας.

Συγκρίνοντας τις τιμές ακρίβειας της πρώτης ομάδας πειραμάτων, όπου ισχύει $treeLength = 1$ με τη δεύτερη και τρίτη ομάδα πειραμάτων, όπου ισχύει $treeLength = 2$, παρατηρείται μεγαλύτερη αποτελεσματικότητα για μικρότερη τιμή ύψους του δένδρου. Αυτό, όπως και προηγουμένως και όπως και στα προηγούμενα πειράματα, οφείλεται στην εξειδίκευση του μοντέλου του αλγορίθμου όσο αυξάνεται το ύψος του δένδρου.

Συγκρίνοντας τις ομάδες πειραμάτων στις οποίες ισχύει $treeLength = 2$, παρατηρείται μεγαλύτερη τιμή ακρίβειας για τιμή παραμέτρου του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών $m = 1$ σε σύγκριση με $m = 2$ για μεγάλο αριθμό κ συνδυασμών $\kappa \in [3,5]$, ενώ για τιμές αριθμού κ συνδυασμών $\kappa \in [1,2]$ παρατηρείται μία υπεροχή της δεύτερης περίπτωσης σε σύγκριση με την πρώτη. Μάλιστα, στη δεύτερη περίπτωση $m = 2$, για $\kappa = 1$ παρατηρείται μία ραγδαία αύξηση, σε σύγκριση με τις άλλες τιμές της τιμής ακρίβειας. Για μεγάλο αριθμό κ συνδυασμών, η δεύτερη ομάδα πειραμάτων έχει μία μικρή υπεροχή σε σύγκριση με την τρίτη ομάδα πειραμάτων όσο αφορά την αποτελεσματικότητα. Αυτό, όπως και στα προηγούμενα πειράματα, οφείλεται στο γεγονός ότι για τιμές $l > 1$ του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών, το μοντέλο εκπαίδευσης είναι πιο γενικό, αφού οι χρονικοί κανόνες συσχέτισης της διαδικασίας ελέγχου έχουν μεταβλητό αριθμό στοιχείων στην κεφαλή τους, ο οποίος ανήκει στο διάστημα $[1, l]$. Για μικρό αριθμό κ συνδυασμών, παρατηρείται υπεροχή της τρίτης ομάδας πειραμάτων σε σύγκριση με τη δεύτερη. Αυτό ερμηνεύεται ότι οφείλεται σε ένα συνδυασμό της τιμής παραμέτρου $m > 1$, η οποία περιορίζει τους προς έλεγχο χρονικούς κανόνες συσχέτισης, αποκόπτοντας από τους κανόνες που υπολογίζονται στη μετρική ακρίβειας τους κανόνες για τους οποίους δεν υπάρχει επαλήθευση για όλα τα στοιχεία του σώματος, και της γενίκευσης του μοντέλου, η οποία επιτυγχάνεται με μείωση του αριθμού κ συνδυασμών.

Είσοδος: $numOfIterations = 8000$, $eventsIterationThreshold = 5$, $startTestingIteration = 100$, $eventDetectionAlgorithm = CUSUM$, $agingFunction = NONE$, $probabilityThreshold = 0.5$.



	CUSUM ALGORITHM		
KFIXEDCOMBINATIONS	**Precision m=1, l=1**	**Precision m=1, l=2**	**Precision m=2, l=1**
5	43.69905054	45.22127142	4.65023901
4	43.83647496	45.01850424	4.717110576
3	46.30505624	44.25655224	5.550085493
2	52.88266342	42.67849149	9.493526788
1	57.53865311	40.86722775	22.86469375

Γραφική παράσταση 10: Πειράματα μεταβολής τιμής σταθερού αριθμού κ συνδυασμών με ανίχνευση συμβάντων με τον αλγόριθμο συσσωρευτικού αθροίσματος

Τα πειράματα εκτελέστηκαν με τον ίδιο αλγόριθμο ανίχνευσης συμβάντων, τον αλγόριθμο συσσωρευτικού αθροίσματος, και έγινε ομαδοποίηση των πειραμάτων για ίδιες τιμές m και l του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών, μεταβάλλοντας την τιμή του σταθερού αριθμού κ συνδυασμών. Οι τιμές του αριθμού κ συνδυασμών ανήκουν στο διάστημα $[1,5]$. Από τα πειράματα, στις δύο από τις τρεις περιπτώσεις ομάδων πειραμάτων παρατηρείται αύξηση τιμής ακρίβειας όσο μειώνεται ο σταθερός αριθμός κ συνδυασμών. Η παρατήρηση αυτή είναι αναμενόμενη, αφού η μείωση του αριθμού κ συνδυασμών καθιστά το μοντέλο εκπαίδευσης του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών πιο γενικό, με αποτέλεσμα να παρατηρείται αύξηση της αποτελεσματικότητας.

Στην περίπτωση που οι παράμετροι του αλγορίθμου ισούνται με $m = 1$ και $l = 2$, παρατηρείται μείωση της τιμής ακρίβειας όσο μειώνεται ο αριθμός των κ συνδυασμών, παρατήρηση η οποία είναι αντίστροφη από το αναμενόμενο, όπως αυτό καθορίζεται από

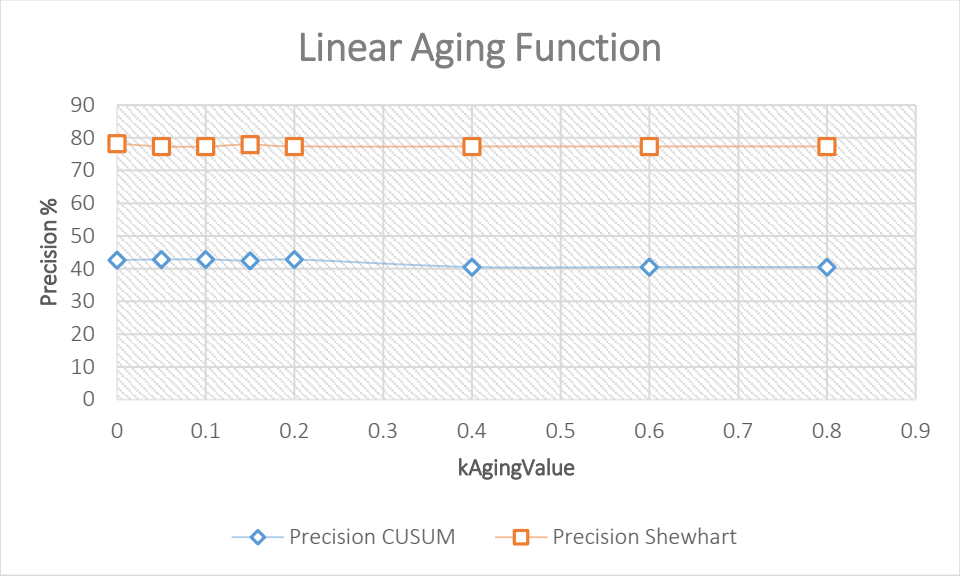
όλα τα υπόλοιπα πειράματα της υποενότητας. Η παρατήρηση αυτή ερμηνεύεται, όπως και στην προηγούμενη υποενότητα, στην υπόθεση που λαμβάνει ο αλγόριθμος συσσωρευτικού αθροίσματος ότι τα δεδομένα εισόδου ακολουθούν κανονική κατανομή, γεγονός που δεν ισχύει για το αρχείο εισόδου της εργασίας. Η υπόθεση του αλγορίθμου συσσωρευτικού αθροίσματος έχει ως αποτέλεσμα εμφάνιση μη αναμενόμενης συμπεριφοράς, όπως μείωση της τιμής ακρίβειας με την αύξηση τιμής κατωφλίου πιθανότητας ή με την μείωση του αριθμού κ συνδυασμών. Με αυτή την υπόθεση ερμηνεύεται και η μη ικανοποιητικές τιμές ακρίβειας στην περίπτωση παραμέτρων αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών $m = 2$ και $l = 1$.

Στην πρώτη περίπτωση, στην οποία οι παράμετροι του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών ισούνται με $m = 1$ και $l = 1$, παρατηρούνται σχετικά ικανοποιητικές τιμές ακρίβειας, οι οποίες αυξάνονται προσεγγιστικά εκθετικά με την μείωση του αριθμού κ συνδυασμών. Αυτό οφείλεται στην γενίκευση του μοντέλου εκπαίδευσης του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών, λόγω του ύψους των δένδρων $treeLength = 1$ στην ιστορική δομή του αλγορίθμου. Με την μείωση του αριθμού κ συνδυασμών, αυξάνεται η γενίκευση του μοντέλου, και συνεπώς αυξάνεται η αποτελεσματικότητα του αλγορίθμου.

7.4.3 Πειράματα μεταβολής τιμής k συνάρτησης απόσβεσης

Τα πειράματα της υποενότητας αυτής εκτελέστηκαν με σταθερές τιμές όλων των παραμέτρων, εκτός από την παράμετρο k της συνάρτησης απόσβεσης, η οποία χρησιμοποιείται στα πλαίσια της διαδικασίας προσαρμοστικού φιλτραρίσματος. Η μελέτη των πειραμάτων ομαδοποιείται για κοινές τιμές των παραμέτρων του συστήματος, και με διαφοροποίηση της συνάρτησης απόσβεσης. Η μελέτη των πειραμάτων γίνεται αρχικά με τη γραμμική συνάρτηση απόσβεσης και στη συνέχεια με την εκθετική συνάρτηση απόσβεσης. Στα αποτελέσματα συμπεριλαμβάνονται οι τιμές ακρίβειας με την μη εκτέλεση της διαδικασίας προσαρμοστικού φιλτραρίσματος, δηλαδή για τιμή k συνάρτησης απόσβεσης $k = 0$.

Είσοδος: $numOfIterations = 8000$, $m = 1$, $l = 1$, $eventsIterationThreshold = 3$, $kFixedCombinations = 0$, $startTestingIteration = 100$, $agingFunction = LINEAR$, $probabilityThreshold = 0.8$



LINEAR AGING FUNCTION		
KAGINGVALUE	**Precision CUSUM**	**Precision Shewhart**
0	42.66304348	78.25156874
0.05	42.81842818	77.34712583
0.1	42.81842818	77.35009165
0.15	42.45014245	77.9435269
0.2	42.81842818	77.35602094
0.4	40.51282051	77.35602094
0.6	40.51282051	77.35602094
0.8	40.51282051	77.35602094

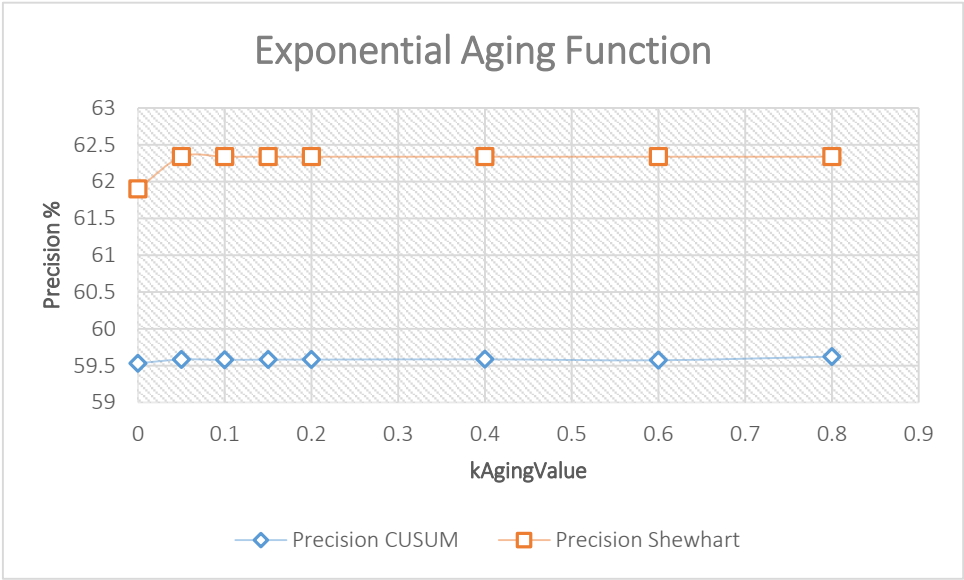
Γραφική παράσταση 11: Πειράματα μεταβολής τιμής k της γραμμικής συνάρτησης απόσβεσης

Σε όλα τα πειράματα της γραφικής παράστασης παρατηρείται υπεροχή του αλγορίθμου διαγραμμάτων ελέγχου Shewhart σε σύγκριση με τον αλγόριθμο συσσωρευτικού αθροίσματος, όσο αφορά τη διαδικασία ανίχνευσης συμβάντων σε ροές δεδομένων αισθητήρων πολλαπλών μεταβλητών. Αυτό οφείλεται στο προσαρμοστικό χαρακτήρα του αλγορίθμου διαγραμμάτων ελέγχου Shewhart, σε σύγκριση με τον αλγόριθμο συσσωρευτικού αθροίσματος. Στην περίπτωση ανίχνευσης συμβάντων με τον αλγόριθμο συσσωρευτικού αθροίσματος, παρατηρείται μία μικρή αύξηση της τιμής ακρίβειας με την αύξηση της τιμής k γραμμικής συνάρτησης απόσβεσης, για τιμές k να ανήκουν στο διάστημα $(0,1]$. Αυτό ερμηνεύεται ως η επίτευξη, έστω και μικρής, αύξησης της

αποτελεσματικότητας χρησιμοποιώντας τη διαδικασία προσαρμοστικού φιλτραρίσματος, και συγκεκριμένα χρησιμοποιώντας τη γραμμική συνάρτηση απόσβεσης. Για τιμές k να ανήκουν στο διάστημα $(0.1,0.8]$ παρατηρείται μία μικρή διακύμανση και σύγκλιση σε τιμή ακρίβειας, η οποία είναι χαμηλότερη από την τιμή ακρίβειας με τη μη εφαρμογή της διαδικασίας προσαρμοστικού φιλτραρίσματος. Αυτό οφείλεται στο γεγονός ότι, για σχετικά μεγάλες τιμές k της συνάρτησης απόσβεσης, η διαδικασία προσαρμοστικού φιλτραρίσματος αναθέτει μεγάλο βάρος στις πολύ πρόσφατες ή στην πιο πρόσφατη τιμή πιθανότητας κάποιου χρονικού κανόνα συσχέτισης. Γενικά, αύξηση της τιμής k συνάρτησης απόσβεσης οδηγεί σε ανάθεση βαρών στις τιμές πιθανότητας κάποιου χρονικού κανόνα συσχέτισης με τέτοιο τρόπο, ώστε οι πιο πρόσφατες τιμές πιθανότητας να είναι πιο χρήσιμες από τις πιο παλιές. Για πολύ μεγάλες τιμές k παρατηρείται σύγκλιση τιμής ακρίβειας, αφού λαμβάνει μεγάλη χρησιμότητα η πιο πρόσφατη τιμή πιθανότητας, σε σύγκριση με τις πιο παλιές.

Στην περίπτωση ανίχνευσης συμβάντων με τον αλγόριθμο διαγραμμάτων ελέγχου Shewhart, παρατηρείται επίσης μία μικρή αύξηση της αποτελεσματικότητας χρησιμοποιώντας τη διαδικασία προσαρμοστικού φιλτραρίσματος. Σε αντίθεση με την ομάδα πειραμάτων με ανίχνευση συμβάντων με τον αλγόριθμο συσσωρευτικού αθροίσματος, η τιμή ακρίβειας είναι για όλες τις τιμές k, οι οποίες ανήκουν στο διάστημα $(0,0.8]$, είναι μεγαλύτερη σε σύγκριση με την τιμή ακρίβειας που προκύπτει από την μη εφαρμογή της διαδικασίας προσαρμοστικού φιλτραρίσματος. Η τιμή σύγκλισης ακρίβειας στην περίπτωση αυτή είναι μεγαλύτερη από την τιμή ακρίβειας με τη μη εφαρμογή της διαδικασίας προσαρμοστικού φιλτραρίσματος. Αυτό οφείλεται στο ότι στην περίπτωση αυτή, ακόμα και με την ανάθεση μεγάλης χρησιμότητας στην πιο πρόσφατη τιμή πιθανότητας σε σύγκριση με τις πιο παλιές τιμές πιθανότητας, η αποτελεσματικότητα είναι σε μικρό βαθμό μεγαλύτερη.

Είσοδος: $numOfIterations = 8000$, $m = 1$, $l = 1$, $eventsIterationThreshold = 3$, $kFixedCombinations = 0$, $startTestingIteration = 100$, $agingFunction = EXPONENTIAL$, $probabilityThreshold = 0.5$.



	EXPONENTIAL AGING FUNCTION	
KAGINGVALUE	Precision CUSUM	Precision Shewhart
0	59.53307393	61.89879876
0.05	59.58448016	62.33802554
0.1	59.57919034	62.33802554
0.15	59.58277852	62.33802554
0.2	59.58277852	62.33802554
0.4	59.58636606	62.33802554
0.6	59.57390146	62.33802554
0.8	59.62274224	62.33802554

Γραφική παράσταση 12: Πειράματα μεταβολής τιμής k της εκθετικής συνάρτησης απόσβεσης

Σε όλα τα πειράματα της γραφικής παράστασης, όπως και στην προηγούμενη γραφική παράσταση, παρατηρείται υπεροχή του αλγορίθμου διαγραμμάτων ελέγχου Shewhart σε σύγκριση με τον αλγόριθμο συσσωρευτικού αθροίσματος, όσο αφορά τη διαδικασία ανίχνευσης συμβάντων σε ροές δεδομένων αισθητήρων πολλαπλών μεταβλητών, λόγω του προσαρμοστικού χαρακτήρα του αλγορίθμου διαγραμμάτων ελέγχου Shewhart. Στην περίπτωση ανίχνευσης συμβάντων με τον αλγόριθμο συσσωρευτικού αθροίσματος, παρατηρείται μικρή αύξηση της τιμής ακρίβειας για όλες τις τιμές k, οι οποίες ανήκουν στο διάστημα $(0,0.8]$. Αυτό ερμηνεύεται ως η μικρή αύξηση της αποτελεσματικότητας με την εφαρμογή της διαδικασίας προσαρμοστικού φιλτραρίσματος με την εκθετική συνάρτηση απόσβεσης. Στην περίπτωση ανίχνευσης συμβάντων με τον αλγόριθμο διαγραμμάτων ελέγχου Shewhart, επίσης παρατηρείται μικρή αύξηση της τιμής ακρίβειας με την εφαρμογή της διαδικασίας προσαρμοστικού φιλτραρίσματος. Επίσης,

παρατηρείται μία σταθερή τιμή ακρίβειας για τιμές k, οι οποίες ανήκουν στο διάστημα $[0.05,0.8]$. Αυτό ερμηνεύεται ως η πιο γρήγορη σύγκλιση τιμής ακρίβειας με τη διαδικασία προσαρμοστικού φιλτραρίσματος με την εκθετική συνάρτηση απόσβεσης, συμπεριφορά η οποία είναι αναμενόμενη. Με άλλα λόγια, η πιο πρόσφατη τιμή πιθανότητας έχει μεγαλύτερη χρησιμότητα από τις πιο παλιές τιμές πιθανότητας για πιο μικρές τιμές k της εκθετικής συνάρτησης απόσβεσης, σε σύγκριση με την περίπτωση προσαρμοστικού φιλτραρίσματος με τη γραμμική συνάρτηση απόσβεσης.

Συγκρίνοντας τις τιμές ακρίβειας με την εφαρμογή προσαρμοστικού φιλτραρίσματος με τη γραμμική συνάρτηση απόσβεσης, σε σύγκριση με την εκθετική συνάρτηση απόσβεσης, παρατηρείται ότι στη δεύτερη περίπτωση για πιο μικρές τιμές k συνάρτησης απόσβεσης υπάρχει σύγκλιση τιμής ακρίβειας. Η σύγκλιση αυτή οφείλεται στην ανάθεση μεγάλης χρησιμότητας στις πιο πρόσφατες τιμές πιθανότητας, ή στην πιο πρόσφατη τιμή πιθανότητας κάποιου χρονικού κανόνα συσχέτισης, σε σύγκριση με τις πιο παλιές τιμές πιθανότητας. Η πιο σύντομη σύγκλιση τιμών ακρίβειας στη δεύτερη περίπτωση οφείλεται στη συμπεριφορά της εκθετικής συνάρτησης απόσβεσης. Χρησιμοποιώντας την εκθετική συνάρτηση απόσβεσης, οι τιμές βαρών που ανατίθενται στις πιο πρόσφατες τιμές πιθανότητας κάποιου χρονικού κανόνα συσχέτισης είναι μεγαλύτερες από τις αντίστοιχες τιμές βαρών της γραμμικής συνάρτησης απόσβεσης. Συνεπώς, στην περίπτωση προσαρμοστικού φιλτραρίσματος με την εκθετική συνάρτηση απόσβεσης, οι πιο πρόσφατες τιμές πιθανότητας αποκτούν μεγάλη χρησιμότητα πιο γρήγορα, σε σύγκριση με την γραμμική συνάρτηση απόσβεσης.

8. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην εργασία αυτή έγινε συζήτηση διαφόρων σταδίων της αλυσίδας επεξεργασίας στο πλαίσιο των δεδομένων αισθητήρων ροής πολλαπλών μεταβλητών. Αρχικά, παρουσιάστηκε ο τρόπος με τον οποίο οι τεχνικές ανίχνευσης μεταβολών μπορούν να χρησιμοποιηθούν για το προσδιορισμό συμβάντων σε δίκτυα αισθητήρων. Στη συνέχεια, έγινε συζήτηση μίας προσέγγισης για το συσχετισμό δεδομένων συμβάντων πολλαπλών μεταβλητών. Ο αλγόριθμος της προσέγγισης αυτής περιλαμβάνει ένα μεταβλητής τάξης μοντέλο για την καταγραφή ακολουθιών πολλαπλών συμβάντων. Επίσης, έγινε παρουσίαση του τρόπου με τον οποίο ένα πιθανοτικό πλαίσιο χρονικής γνώσης μπορεί να περιλαμβάνει την αντιπροσώπευση των εξαρτήσεων, με σκοπό την πρόβλεψη μελλοντικών ακολουθιών. Το πλαίσιο αυτό επιτρέπει επίσης τη διατύπωση κανόνων. Για την αντιμετώπιση του προβλήματος που προκύπτει από τις παρωχημένες εξαρτήσεις, έγινε μελέτη ενός χρονικά εξαρτώμενου πλαισίου, το οποίο φιλτράρει τους προκύπτοντες κανόνες με την πάροδο του χρόνου μέσα από παράγοντες απόσβεσης. Η πειραματική αξιολόγηση των προσεγγίσεων που παρουσιάστηκαν βασίζεται σε δεδομένα πραγματικού κόσμου, τα οποία προέρχονται από το τομέα της ναυτιλίας.

Στα πειράματα που εκτελέστηκαν στα πλαίσια της εργασίας παρατηρήθηκε υπεροχή του αλγορίθμου διαγραμμάτων ελέγχου Shewhart, σε σύγκριση με τον αλγόριθμο συσσωρευτικού αθροίσματος, όσο αφορά την ανίχνευση συμβάντων σε ροές δεδομένων αισθητήρων πολλαπλών μεταβλητών. Η υπεροχή καθορίζεται από τις τιμές ακρίβειας, οι οποίες στις περισσότερες περιπτώσεις ήταν μεγαλύτερες για ίδιες τιμές όλων των παραμέτρων, αλλά και από τη συμπεριφορά των τιμών ακρίβειας με αύξηση ή μείωση κάποιας παραμέτρου και σύγκριση με την αναμενόμενη συμπεριφορά. Η υπεροχή του αλγορίθμου διαγραμμάτων ελέγχου Shewhart ερμηνεύεται να οφείλεται στην προσαρμοστική του συμπεριφορά. Ο αλγόριθμος συσσωρευτικού αθροίσματος λαμβάνει υπόψη ότι τα δεδομένα εισόδου ακολουθούν κανονική κατανομή, γεγονός που δεν ισχύει για τα δεδομένα εισόδου των πειραμάτων.

Η αύξηση της τιμής κατωφλίου πιθανότητας των χρονικών κανόνων συσχέτισης, οι οποίοι συμπεριλαμβάνονται στη διαδικασία ελέγχου, οδηγεί σε αύξηση της αποτελεσματικότητας του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών. Η αύξηση του ύψους των δένδρων στην ιστορική δομή του αλγορίθμου καθιστά το μοντέλο εκπαίδευσης του αλγορίθμου πιο συγκεκριμένο, με αποτέλεσμα να υπάρχει μείωση της αποτελεσματικότητας. Όσο αυξάνεται η τιμή της

παραμέτρου m του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών, τόσο αυξάνονται τα κριτήρια ένταξης των χρονικών κανόνων συσχέτισης στη διαδικασία ελέγχου, με αποτέλεσμα την αύξηση των τιμών ακρίβειας. Επίσης, όσο αυξάνεται η τιμή της παραμέτρου l του αλγορίθμου, τόσο αυξάνεται ο αριθμός των χρονικών κανόνων συσχέτισης, οι οποίοι συμπεριλαμβάνονται στη διαδικασία ελέγχου. Αυτό οφείλεται στην ποικιλομορφία του αριθμού στοιχείων στην κεφαλή των χρονικών κανόνων συσχέτισης, ο οποίος ανήκει στο διάστημα $[1, l]$.

Η μείωση της τιμής σταθερού αριθμού κ συνδυασμών οδηγεί σε γενίκευση του μοντέλου εκπαίδευσης του αλγορίθμου μεταβλητής τάξης συσχέτισης δεδομένων συμβάντων πολλαπλών μεταβλητών. Η γενίκευση του μοντέλου οφείλεται στη μείωση του μέγιστου αριθμού συμβάντων που μπορούν να συμπεριληφθούν σε κάποιο κόμβο ενός δένδρου στην ιστορική δομή του αλγορίθμου, και συνεπώς στην απλούστευση της μορφής των χρονικών κανόνων συσχέτισης. Η γενίκευση που προκύπτει από τη μείωση του σταθερού αριθμού κ συνδυασμών οδηγεί σε αύξηση της αποτελεσματικότητας.

Η εφαρμογή της διαδικασίας προσαρμοστικού φιλτραρίσματος οδηγεί σε μικρή αύξηση της τιμής αποτελεσματικότητας, η οποία οφείλεται στην ανάθεση μεγαλύτερης χρησιμότητας στις πιο πρόσφατες τιμές πιθανότητας κάποιου χρονικού κανόνα συσχέτισης, σε σύγκριση με τις πιο παλιές τιμές πιθανότητας. Για κάποια οριακή τιμή της παραμέτρου k της συνάρτησης απόσβεσης, η οποία αποτελεί την προσέγγιση προσαρμοστικού φιλτραρίσματος της εργασίας, παρατηρείται σύγκλιση των τιμών ακρίβειας. Για μεγαλύτερες τιμές της παραμέτρου k, η αποτελεσματικότητα παραμένει αμετάβλητη. Η σύγκλιση αυτή οφείλεται στην ανάθεση μεγάλης χρησιμότητας στις πιο πρόσφατες ή στην πιο πρόσφατη τιμή πιθανότητας, σε σύγκριση με τις πιο παλιές τιμές πιθανότητας κάποιου χρονικού κανόνα συσχέτισης. Η σύγκλιση είναι πιο γρήγορη όταν γίνεται εφαρμογή της εκθετικής συνάρτησης απόσβεσης, σε σύγκριση με τη γραμμική συνάρτηση απόσβεσης.

Στοιχεία μίας κακής ή δυσλειτουργικής συμπεριφοράς ενός δικτύου αισθητήρων είναι συχνά ενσωματωμένα μέσα σε ακολουθίες συμβάντων, οι οποίες ανιχνεύθηκαν μέσω κατανεμημένων κόμβων σε όλο το δίκτυο. Μία τέτοια κατάσταση μπορεί να απαιτεί τη επεξεργασία με βάση πολλαπλά χρονικά βήματα, έτσι ώστε να είναι εμφανή τα αποτελέσματα της. Επίσης, αρκετές καταστάσεις οι οποίες μπορούν να χαρακτηριστούν κακές, μπορεί να οδηγήσουν σε παρόμοιου τύπου συμβάντα. Μελλοντική έρευνα θα μπορούσε να επικεντρωθεί στην αναγνώριση κρυμμένων ή μη παρατηρήσιμων καταστάσεων του συστήματος, οι οποίες μπορούν να επηρεάσουν τις παρατηρήσιμες

μεταβλητές ή τη συμπεριφορά των αισθητήρων. Τα κρυμμένα μαρκοβιανά μοντέλα είναι κατάλληλα για τη δημιουργία ενός πλαισίου μοντελοποίησης, για την αναπαράσταση τέτοιων σχέσεων αιτιότητας και το καθορισμό των μη παρατηρήσιμων καταστάσεων ενός συστήματος σε σχέση με τις παρατηρούμενες μετρήσεις.

Μία άλλη πτυχή μεγάλης σημασίας είναι ένας πιο βελτιωμένος χαρακτηρισμός των ανιχνευόμενων μεταβολών στο πλαίσιο μιας χρονοσειράς. Οι περισσότεροι των υπαρχόντων αλγορίθμων ανίχνευσης μεταβολών δεν παρέχουν επιπλέον σχολιασμό για τα ανιχνευόμενα συμβάντα. Με άλλα λόγια, ο χαρακτηρισμός είναι μία απλή δυαδική συνάρτηση εμφάνισης ή όχι κάποιου συμβάντος. Ως ένα επόμενο βήμα, θα μπορούσε να γίνει κάποια επέκταση των υπάρχοντων τεχνικών ανίχνευσης μεταβολών μίας μεταβλητής, έτσι ώστε να μπορούν να παρέχουν ανίχνευση μεταβολών για αριθμητικά χαρακτηριστικά, ανίχνευση η οποία να είναι πολλαπλών επιπέδων. Το βήμα αυτό περιλαμβάνει το προσδιορισμό νέου εύρους τιμών με νέες τιμές εισόδου, οι οποίες δεν είναι σύμφωνες με τις υπάρχουσες ομάδες τιμών.

ΑΝΑΦΟΡΕΣ

[1] L. Fan, P. Cao, W. Lin και Q. Jacobson, «Web prefetching between low-bandwidth clients and proxies: potential and performance,» *ACM SIGMETRICS Performance Evaluation Review,* τόμ. 27, αρ. 1, pp. 178-187, 1999.

[2] A. Harutyunyan, P. A. C. U. VMware, A. Poghosyan, N. Grigoryan και M. Marvasti, «Abnormality analysis of streamed log data,» σε *Network Operations and Management Symposium (NOMS)*, Krakow, 2014.

[3] M. A. Marvasti, A. V. Poghosyan, A. N. Harutyunyan και N. M. Grigoryan, «An anomaly event correlation engine: Identifying root causes, bottlenecks, and black swans in IT environments,» *VMware Technical Journal,* τόμ. 2, αρ. 1, pp. 35-45, 2013.

[4] L. I. Kuncheva, «Change detection in streaming multivariate data using likelihood detectors,» *Knowledge and Data Engineering, IEEE Transactions on,* τόμ. 25, αρ. 5, pp. 1175-1180, 2013.

[5] T. Dasu, S. Krishnan, S. Venkatasubramanian και K. Yi, «An information-theoretic approach to detecting changes in multi-dimensional data streams,» σε *In Proc. Symp. on the Interface of Statistics, Computing Science, and Applications*, Citeseer, 2006.

[6] H. Hotelling, «The Generalization of Student's Ratio,» *The Annals of Mathematical Statistics,* τόμ. 2, αρ. 3, pp. 360-378, 1931.

[7] G. Jiang και G. Cybenko, «Temporal and spatial distributed event correlation for network security,» σε *American Control Conference, 2004. Proceedings of the 2004*, 2004.

[8] O. Cappe, E. Moulines και T. Ryden, «Inference in Hidden Markov Models,» 2005.

[9] R. E. Kalman, «A new approach to linear filtering and prediction problems,» *Journal of Fluids Engineering,* τόμ. 82, αρ. 1, pp. 35-45, 1960.

[10] M. Severo και J. Gama, «Change detection with kalman filter and cusum,» σε *Discovery Science*, Springer, 2006, pp. 243-254.

[11] E. Page, «Continuous inspection schemes,» *Biometrika,* pp. 100-115, 1954.

[12] J. Veldman, H. Wortmann και W. Klingenberg, «Typology of condition based maintenance,» *Journal of Quality in Maintenance Engineering,* τόμ. 17, αρ. 2, pp. 183-202, 2011.

[13] M. Basseville και I. V. Nikiforov, Detection of abrupt changes: theory and application, τόμ. 104, Prentice Hall Englewood Cliffs, 1993.

[14] S. Alestra, C. Bordry, C. Brand, E. Burnaev, P. Erofeev, A. Papanov και C. Silveira-Freixo, «Rare event anticipation and degradation trending for aircraft predictive maintenance,» *11th World Congress on Computational Mechanics (WCCM XI).*

[15] I. Steinwart και A. Christmann, Support vector machines, Springer Science & Business Media, 2008.

[16] A. Marjanovic, G. Kvascev, P. Tadic και Z. Durovic, «Applications of predictive maintenance techniques in industrial systems,» *Serbian Journal of Electrical Engineering,* τόμ. 8, αρ. 3, pp. 263-279, 2011.

[17] K. Fukunaga, Introduction to statistical pattern recognition, Academic press, 2013.

[18] A. J. Viterbi, «Error bounds for convolutional codes and an asymptotically optimum decoding algorithm,» *Information Theory, IEEE Transactions on,* τόμ. 13, αρ. 2, pp. 260-269, 1967.

[19] L. R. Rabiner και B. Gold, «Theory and application of digital signal processing,» *Englewood Cliffs, NJ, Prentice-Hall, Inc., 1975. 777 p.,* τόμ. 1, 1975.

[20] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam και E. Cayirci, «A survey on sensor networks,» *Communications magazine, IEEE,* τόμ. 40, αρ. 8, pp. 102-114, 2002.

[21] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam και E. Cayirci, «Wireless sensor networks: a survey,» *Computer networks,* τόμ. 38, αρ. 4, pp. 393-422, 2002.

[22] S. Rajasegarar, C. Leckie και M. Palaniswami, «Anomaly detection in wireless sensor networks,» *Wireless Communications, IEEE,* τόμ. 15, αρ. 4, pp. 34-40, 2008.

[23] V. Chandola, A. Banerjee και V. Kumar, «Anomaly detection: A survey,» *ACM computing surveys (CSUR),* τόμ. 41, αρ. 3, p. 15, 2009.

[24] Y. Zhang, N. Meratnia και P. Havinga, «Outlier detection techniques for wireless sensor networks: A survey,» *Communications Surveys & Tutorials, IEEE,* τόμ. 12, αρ. 2, pp. 159-170, 2010.

[25] D. C. Montgomery, Introduction to Statistical Quality Control, 2005.

[26] H. Luetkepohl, Introduction to multiple time series analysis, Springer, 1991.

[27] D. B. Neill και G. F. Cooper, «A multivariate Bayesian scan statistic for early event detection and characterization,» *Machine learning,* τόμ. 79, αρ. 3, pp. 261-282, 2010.

[28] B. Krishnamachari και S. Iyengar, «Distributed Bayesian algorithms for fault-tolerant event region detection in wireless sensor networks,» *Computers, IEEE Transactions on,* τόμ. 53, αρ. 3, pp. 241-250, 2004.

[29] C. C. Aggarwal, Managing and mining sensor data, Springer Science & Business Media, 2013.

[30] J. Han, M. Kamber και J. Pei, Data mining: concepts and techniques: concepts and techniques, Elsevier, 2011.

[31] B. Scholkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola και R. C. Williamson, «Estimating the support of a high-dimensional distribution,» *Neural computation,* τόμ. 13, αρ. 7, pp. 1443-1471, 2001.

[32] M. Frisen, «Properties and use of the Shewhart method and its followers,» *Sequential Analysis,* τόμ. 26, αρ. 2, pp. 171-193, 2007.

[33] J. H. Thomas, «Statistical process control procedures for correlated observations,» *The Canadian Journal of Chemical Engineering,* τόμ. 69, 1991.

[34] F. Gustafsson και F. Gustafsson, Adaptive filtering and change detection, τόμ. 1, Wiley New York, 2000.

[35] W. H. Woodall, «The distribution of the run length of one-sided CUSUM procedures for continuous random variables,» *Technometrics,* τόμ. 25, αρ. 3, pp. 295-301, 1983.

[36] F. Alt και K. Jain, «Multivariate quality controlMultivariate quality control,» σε *Encyclopedia of Operations Research and Management Science*, Springer, 2001, pp. 544-550.

[37] L. A. Jones και W. H. Woodall, «The performance of bootstrap control charts,» *Journal of Quality Technology,* τόμ. 30, αρ. 4, p. 362, 1998.

[38] E. Zivot και J. Wang, «Vector autoregressive models for multivariate time series,» *Modeling Financial Time Series with S-PLUS,* pp. 385-429, 2006.

[39] A. Neumaier και T. Schneider, «Estimation of parameters and eigenmodes of multivariate autoregressive models,» *ACM Transactions on Mathematical Software (TOMS),* τόμ. 27, αρ. 1, pp. 27-57, 2001.

[40] J. H. Stock και M. W. Watson, «Vector autoregressions,» *Journal of Economic perspectives,* pp. 101-115, 2001.

[41] S. Roberts, «Control chart tests based on geometric moving averages,» *Technometrics,* τόμ. 1, αρ. 3, pp. 239-250, 1959.

[42] J. M. Lucas και M. S. Saccucci, «Exponentially weighted moving average control schemes: properties and enhancements,» *Technometrics,* τόμ. 32, αρ. 1, pp. 1-12, 1990.

[43] L. Fan, P. Cao, W. Lin και Q. Jacobson, «Web prefetching between low-bandwidth clients and proxies: potential and performance,» σε *ACM SIGMETRICS Performance Evaluation Review*, τόμ. 27, ACM, 1999, pp. 178-187.

[44] B. Knoll, «Ensemble Prediction by Partial Matching,» *Computer Science,* τόμ. 540.

[45] A. Moffat, «Implementing the PPM data compression scheme,» *Communications, IEEE Transactions on,* τόμ. 38, αρ. 11, pp. 1917-1921, 1990.

[46] T. Palpanas και A. Mendelzon, Web prefetching using partial match prediction, Citeseer, 1998.

[47] J. Wang, «A survey of web caching schemes for the internet,» *ACM SIGCOMM Computer Communication Review,* τόμ. 29, αρ. 5, pp. 36-46, 1999.

[48] K. Kapitanova, S. H. Son και K.-D. Kang, «Using fuzzy logic for robust event detection in wireless sensor networks,» *Ad Hoc Networks,* τόμ. 10, αρ. 4, pp. 709-722, 2012.

[49] E. Goetz και S. Shenoi, Critical infrastructure protection, Springer Heidelberg, 2008.

[50] P. Ni, L. Wan και Y. Cai, «Event Correlations in Sensor Networks,» σε *Computational Science--ICCS 2009*, Springer, 2009, pp. 500-509.

[51] I. F. Akyildiz, M. C. Vuran και O. B. Akan, «On exploiting spatial and temporal correlation in wireless sensor networks,» σε *Proceedings of WiOpt*, τόμ. 4, 2004, pp. 71-80.

[52] M. Stamp, «A revealing introduction to hidden Markov models,» *Department of Computer Science San Jose State University,* 2004.

[53] D. L. Isaacson και R. W. Madsen, Markov chains, theory and applications, τόμ. 4, Wiley New York, 1976.

[54] R. Begleiter, R. El-Yaniv και G. Yona, «On prediction using variable order Markov models,» *Journal of Artificial Intelligence Research,* pp. 385-421, 2004.

[55] G. Bejerano, «Algorithms for variable length Markov chain modeling,» *Bioinformatics,* τόμ. 20, αρ. 5, pp. 788-789, 2004.

[56] K. Gerdes, E. Hajicova και L. Wanner, Computational Dependency Theory, τόμ. 258, IOS Press, 2013.

[57] M.-C. De Marneffe, T. Dozat, N. Silveira, K. Haverinen, F. Ginter, J. Nivre και C. D. Manning, «Universal Stanford Dependencies: A cross-linguistic typology,» σε *Proceedings of LREC*, 2014.

[58] S. Horwitz, T. Reps και D. Binkley, «Interprocedural slicing using dependence graphs,» *ACM Transactions on Programming Languages and Systems (TOPLAS),* τόμ. 12, αρ. 1, pp. 26-60, 1990.

[59] R. Bellman, «On a routing problem,» 1956.

[60] L. Ford και D. R. Fulkerson, «Flows in networks,» *Princeton Princeton University Press,* 1962.

[61] E. F. Moore, «The shortest path through a maze,» 1959.

[62] E. W. Dijkstra, «A note on two problems in connexion with graphs,» *Numerische mathematik,* τόμ. 1, αρ. 1, pp. 269-271, 1959.

[63] T. Nakagawa, K. Inui και S. Kurohashi, «Dependency tree-based sentiment classification using CRFs with hidden variables,» σε *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, 2010, pp. 786-794.

[64] D. Gamerman, Markov Chain Monte Carlo: stochastic simulation for bayesian inference, London: Chapman & Hall, 1997.

[65] L. De Raedt, A. Kimmig και H. Toivonen, «ProbLog: A Probabilistic Prolog and Its Application in Link Discovery,» σε *IJCAI*, τόμ. 7, 2007, pp. 2462-2467.

[66] D. Fierens, G. Van den Broeck, J. Renkens, D. Shterionov, B. Gutmann, I. Thon, G. Janssens και L. De Raedt, «Inference and learning in probabilistic logic programs using weighted Boolean formulas,» *arXiv preprint arXiv:1304.6810,* 2013.

[67] P. Shakarian, A. Parker, G. Simari και V. V. Subrahmanian, «Annotated probabilistic temporal logic,» *ACM Transactions on Computational Logic (TOCL),* τόμ. 12, αρ. 2, p. 14, 2011.

[68] A. Dekhtyar, M. I. Dekhtyar και V. Subrahmanian, «Temporal Probabilistic Logic Programs.,» σε *ICLP*, τόμ. 99, 1999, pp. 109-123.

[69] J. P. Dickerson, G. I. Simari και V. Subrahmanian, «Using Temporal Probabilistic Rules to Learn Group Behavior,» *Handbook of Computational Approaches to Counterterrorism,* p. 245, 2012.

[70] R. Agrawal και R. Srikant, «Fast algorithms for mining association rules,» σε *Proc. 20th int. conf. very large data bases, VLDB*, 1994.

[71] H. Toivonen, «Sampling large databases for association rules,» σε *VLDB*, τόμ. 96, 1996, pp. 134-145.

[72] R. Srikant και R. Agrawal, «Mining quantitative association rules in large relational tables,» σε *ACM SIGMOD Record*, τόμ. 25, ACM, 1996, pp. 1-12.

[73] J. M. Ale και G. H. Rossi, «An approach to discovering temporal association rules,» σε *Proceedings of the 2000 ACM symposium on Applied computing-Volume 1*, 2000.

[74] X. Chen και I. Petrounias, «Discovering temporal association rules: Algorithms, language and system,» σε *icde*, IEEE, 2000, p. 306.

[75] C. P. Rainsford και J. F. Roddick, «Adding temporal semantics to association rules,» σε *Principles of Data Mining and Knowledge Discovery*, Springer, 1999, pp. 504-509.

[76] Y. Li, P. Ning, X. S. Wang και S. Jajodia, «Discovering calendar-based temporal association rules,» *Data & Knowledge Engineering,* τόμ. 44, αρ. 2, pp. 193-218, 2003.

[77] S. Laxman και P. S. Sastry, «A survey of temporal data mining,» *Sadhana,* τόμ. 31, αρ. 2, pp. 173-198, 2006.

[78] B. Babcock, S. Babu, M. Datar, R. Motwani και J. Widom, «Models and issues in data stream systems,» σε *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2002.

[79] A. H. Sayed, Fundamentals of adaptive filtering, John Wiley & Sons, 2003.

[80] R. Mehra, «Approaches to adaptive filtering,» *1970 IEEE Symposium on Adaptive Processes (9th) Decision and Control,* αρ. 9, p. 141, 1970.

[81] P. S. Diniz, Adaptive filtering, Springer, 1997.

[82] I. Koychev, «Gradual forgetting for adaptation to concept drift,» σε *Proceedings of ECAI 2000 Workshop on Current Issues in Spatio-Temporal Reasoning*, 2000.

[83] R. Klinkenberg, «Learning drifting concepts: Example selection vs. example weighting,» *Intelligent Data Analysis,* τόμ. 8, αρ. 3, pp. 281-300, 2004.

[84] J. Gosling, The Java language specification, Addison-Wesley Professional, 2000.

[85] T. Lindholm, F. Yellin, G. Bracha και A. Buckley, The Java virtual machine specification, Pearson Education, 2014.