*ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ*
*ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ*

# ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

# Τεχνικές Μηχανικής Μάθησης στα Διαγράμματα Ελέγχου για Συσχετισμένα Δεδομένα

Ευτυχία Ηλιοπούλου

ΑΜ: 09117046

**Επιβλέπων καθηγητής**: Χρήστος Κουκουβίνος

Καθηγητής ΕΜΠ

Αθήνα,2023

*NATIONAL TECHNICAL UNIVERSITY OF ATHENS*
*SCHOOL OF MATHEMATICAL AND PHYSICAL SCIENCES*

# THESIS

# Machine Learning Control Charts For Correlated Data

Eftihia Iliopoulou

Registration Number: 09117046

**Supervisor Professor**:

Christos Koukoubivos, NTUA Professor

Athens,2023

# Περίληψη

Η εξέλιξη της τεχνολογίας και η ανάπτυξη του κλάδου της Τεχνητής Νοημοσύνης σε συνδιασμό με την ανάγκη επεξεργασίας και ανάλυσης μεγάλων και εξαρτημένων δεδομένων οδήγησε στην στροφή του Στατιστικού Ελέγχου Ποιότητας προς τις Τεχνικές Μηχανικής Μάθησης. Οι Τεχνικές Μηχανικής Μάθησης ενσωματώνονται όλο και περισσότερο στον Στατιστικό Έλεγχο Διεργασιών (ΣΕΔ). Τα διαγράμματα ελέγχου προϋποθέτουν την ανεξαρτησία των δεδομένων, κάτι το οποίο είναι σχεδόν αδύνατο στην εποχή μας, με την πληθώρα των πληροφοριών και των πολυπλοκότερων δεδομένων. Η Μηχανική Μάθηση δίνει την λύση σε αυτά τα προβλήματα. Τεχνικές της χρησιμοποιούνται όλο και περισσότερο είτε σε ήδη υπάρχοντα διαγράμματα είτε σε νέα, που στηρίζονται εξ ολοκλήρου στην Τεχνητή Νοημοσύνη.

Στην παρούσα εργασία εισάγουμε τα διάφορα είδη Μηχανικής Μάθησης που χρησιμοποιούνται στα διαγράμματα ελέγχου και τα χαρακτηριστικά τους καθώς επίσης παρουσιάζονται αναλυτικά ποικίλα διαγράμματα που βασίζονται σε πολλούς διαφορετικούς κλάδους της Μηχανικής Μάθησης. Επιπλέον, αναφέρουμε μια διαδικασία αποσυσχέτισης των δεδομένων που μπορεί να χρησιμοποιηθεί σε συνδιασμό με τα διαγράμματα ελέγχου. Η εργασία ολοκληρώνεται με ένα παράδειγμα πραγματικών δεδομένων στο οποίο εφαρμόζουμε στο πραγματικό σύνολο δεδομένων τρία διαφορετικά διαγράμματα με τεχνικές Μηχανικής Μάθησης και συγκρίνουμε τα αποτελέσματά τους.

***Λέξεις κλειδιά:*** Στατιστικός Έλεγχος Διεργασιών, Τεχνικές Μηχανικής Μάθησης, διαγράμματα ελέγχου, συσχετισμένα δεδομένα, μέθοδος πυρήνα.

# Abstract

The evolution of technology and the development of the field of Artificial Intelligence combined with the need to process and analyze large and correlated data led to the shift of Statistical Quality Control to Machine Learning Techniques. Machine Learning Techniques are increasingly integrated into Statistical Process Control (SPC). Control charts assume independence of data, which is almost impossible in the era of information overload and the era of big data. Machine Learning provides the solution to these problems. Its techniques are being used more and more often either in existing control charts or in new ones, based entirely on Artificial Intelligence.

In this paper we introduce the different techniques of Machine Learning used in control charts and their characteristics as well as details of various charts based on many different Machine Learning techniques. Additionally, we include a data decorrelation procedure that can be used in conjunction with control charts. The paper concludes with a real data example, in which we apply to the real data set three different control charts with Machine Learning techniques and compare the results.

***Keywords:*** Statistical Process Control, Machine Learning Techniques, control charts, correlated data, kernel method.

# Ευχαριστίες

Θα ήθελα να ευχαριστήσω πρωτίστως τον επιβλέποντα καθηγητή της διπλωματικής μου εργασίας, τον Κ. Χρήστο Κουκουβίνο για την καθοδήγηση και την στήριξη του, καθώς και την υποψήφια διδάκτορα του Εθνικού Μετσόβιου Πολυτεχνείου, Αγγελική Λάπα για όλη την βοήθειά της.

Επιπλέον, ευχαριστώ θερμά τους γονείς μου, την αδερφή μου, Άννα Ηλιοπούλου, τους φίλους και συμφοιτητές μου για την υποστήριξη τους όλα τα χρόνια των σπουδών μου, που με έκαναν να πιστέψω στον εαυτό μου και να παλέψω για το καλύτερο.

*Vires acquirit eundo*

# Περιεχόμενα

# Κατάλογος Συντομογραφιών

| | |
|---|---|
| ΣΕΔ | Στατιστικός Έλεγχος Ποιότητας |
| SPC | Statistical Process Control |
| SQC | Statistical Quality Control |
| ML | Machine Learning |
| AI | Artificial Intellingence |
| IDD | Independent and Identically Distributed |
| UCL | Uper Control Limit |
| LCL | Lower Control Limit |
| CL | Ccentral Line |
| NN | Neural Networks |
| BPNN | Back Propagation Neural Network |
| RNN | (Recurrent Neural Network |
| DL | Deep Learning |
| RF | Random Forest |
| DT | Decision Trees |
| IC | In-Control |
| OOC | Out-Of-Control |
| ARL | Average Run Length |
| AC | Artificial Contrast |
| RTC | Real Time Contrast |
| SVM | Support Vector Machine |
| KNN | K Nearest Neighbor |
| NB-k | Naive Bayes kernel |
| DSVM | Distance Support Vector Machine |
| SVDD | Support Vector Data Description |
| KD | Kernel Distance |
| RSVM | Robust Support Vector Machine |
| KPCA | Kernel Principal Component Analysis Control Chart |
| RKNN | Random K Nearest Neighbor |

# Κατάλογος Σχημάτων

# Κατάλογος Πινάκων

# Εισαγωγή

Η χρήση στατιστικών τεχνικών για τον έλεγχο μιας διεργασίας ή μιας μεθόδου παραγωγής ορίζεται ως στατιστικός έλεγχος διεργασιών (ΣΕΔ) (Statistical Process Control - SPC).Τα εργαλεία και οι διαδικασίες SPC μπορούν να βοηθήσουν στην παρακολούθηση της συμπεριφοράς της διαδικασίας, στην ανακάλυψη προβλημάτων σε εσωτερικά συστήματα και στην εύρεση λύσεων για ζητήματα παραγωγής. Ο ΣΕΔ χρησιμοποιείται συχνά εναλλακτικά με τον στατιστικό έλεγχο ποιότητας (Statistical Quality Control - SQC). Στην βιομηχανία,τα διαγράμματα ελέγχου (Control Charts) είναι αποτελεσματικά εργαλεία του SPC για τη συνεχή παρακολούθηση μιας διαδικασίας καθώς και για τον εντοπισμό ανωμαλιών και βελτιστοποίηση της διαδικασίας ()Montgomery D. C. (2013).. Οι ανωμαλίες αυτές των αποτελεσμάτων μπορεί να οφείλονται είτε στην αναπόφευκτη μεταβλητότητα του συστήματος, είτε σε μια σείρα από εξωτερικούς παράγοντες όπως οι αλλαγές στις περιβαλλοντικές συνθήκες, η χρήση των μηχανημάτων, λάθος αποφάσεις του εργατικού δυναμικού, κάποιο ατύχημα κ.τ.λ.. Στα συστήματα που επηρεάζονται από αυτές τις αλλαγές θα υπάρχουν μετατοπίσεις που μαρτυρούν πως κάποιος παράγοντας άλλαξε τις συνθήκες του συστήματος με την πάροδο του χρόνου. Έτσι, τα δεδομένα που λαμβάνουμε είναι πλέον εξαρτημένα (συσχετισμένα - correlated). Τα διαγράμματα ελέγχου του SPC που χρησιμοποιούνται ευρέως είναι τα $\bar{x}$, R, S και $S^2$. Κατά την εφαρμογή των διαγραμμάτων αυτών, γίνεται η υπόθεση πως δεν υπάρχει συσχέτιση μεταξύ των δειγμάτων. Στην πράξη όμως, σε πολλές περιπτώσεις τα δεδομένα είναι συσχετισμένα. Η εργασία του Neuhardt(1987) αμφισβητεί την τυπική υπόθεση ότι τα δείγματα που χρησιμοποιούνται για τη δημιουργία στατιστικών διαγραμμάτων ελέγχου διεργασιών είναι ανεξάρτητα και επεσήμανε ότι όταν υπάρχει θετική συσχέτιση μεταξύ των δειγμάτων, τα όρια ελέγχου για το $\bar{x}$ διάγραμμα θα είναι σημαντικά μεγαλύτερα. Αυτό οδηγεί σε συχνότερη παρατήρηση σημείων εκτός των όριων και υποεκτίμηση της τυπικής απόκλισης του μέσου όρου του δείγματος που προκαλεί υπερεκτίμηση της ικανότητα της διαδικασίας.(Yang K., Hancock W. M. (1990))

Αυτού του είδους τα δεδομένα είναι άφθονα σε πεδία όπως τα οικονομικά, τη βιολογία, την ιατρική, τις φυσικές επιστήμες και την μηχανική, όπου υπάρχει ενδιαφέρον για την κατανόηση των μηχανισμών που διέπουν αυτά τα δεδομένα, την παραγωγή προβλέψεων μελλοντικής συμπεριφοράς και την εξαγωγή συμπερασμάτων από τα δεδομένα.

Η εποχή στην οποία ζούμε χαρακτηρίζεται ως η εποχή των Μεγάλων Δεδομένων (Big Data). Τα δεδομένα που καλούμαστε να επεξεργαστούμε είναι πολυπλοκότερα και προφανώς σε πολύ μεγάλο βαθμό εξαρτημένα. Ήταν απαραίτητο επομένως να εμπλουτίσουμε και να εξελίξουμε τα διαγράμματα ελέγχου με νέες τεχνικές που θα χειρίζονται με επιτυχία τα δεδομένα της εποχής. Καθώς λοιπόν η τεχνολογία εξελίσσεται και τα δεδομένα μας είναι περισσότερα και πολυπλοκότερα οι τεχνικές Machine Learning χρησιμοποιούνται όλο και περισσότερο στον ΣΕΔ (Tran, P.H., Ahmadi Nadi, A., Nguyen, T.H., Tran, K.D., Tran, K.P. (2022).) Η Μηχανική Μάθηση είναι ένας τομέας της Τεχνητής Νοημοσύνης (Artificial Intellingence - AI), ο οποίος αποτελείται από προγραμματιστικούς αλγορίθμους που μαθαίνουν αυτόματα από δεδομένα και εμπειρίες ή μέσω αλληλεπίδρασης με το περιβάλλον. Αυτό που κάνει την Μηχανική Μάθηση πραγματικά χρήσιμη είναι το εξής γεγονός: ο αλγόριθμος μπορεί να μάθει και να προσαρμόσει τα αποτελέσματά του με βάση νέα δεδομένα χωρίς εκ των προτέρων προγραμματισμό. Οι αλγόριθμοι αυτοί ενσωματώνονται στα διαγράμματα ελέγχου και δημιουργούνται νέα και βελτιωμένα διαγράμματα. Η εργασία αυτή αποτελείται από 5 κεφάλαια. Στα Κεφάλαια 1 και 2 κάνουμε μια εισαγωγή στην έννοια των χρονοσειρών και στον πολυμεταβλητό έλεγχο ποιότητας, στο Κεφάλαιο 3 μια παρουσίαση των καταλληλότερων εξισώσεων που θα χρησιμοποιηθούν για τον προσδιορισμό των ορίων ελέγχου στα διαγράμματα ελέγχου για εξαρτημένα δεδομένα, στα κεφάλαια 4 και 5 γίνεται η εισαγωγή των τεχνικών Μηχανικής Μάθησης στον ΣΕΔ και παρατίθονται σχετικά διαγράμματα, και στο κεφάλαιο 6 αναφέρεται ένα παράδειγμα πάνω σε πραγματικά δεδομένα στο οποίο συγκρίνουμε διαφορετικά διαγραμμάτων ελέγχου μηχανικής μάθησης.

# Κεφάλαιο 1

# Χρονοσειρές

Μια χρονοσειρά μπορεί να θεωρηθεί ως μια συλλογή παρατηρήσεων που πραγματοποιούνται διαδοχικά στο χρόνο. Υπάρχουν σειρές που είναι ντετερμινιστικές αλλά και σειρές που συμπεριφέρονται σύμφωνα με τους νόμους των πιθανοτήτων. Ιδιαίτερο ενδιαφέρον παρουσιάζουν οι δεύτερες και σε αυτές θα επικεντρωθούμε. Σε αυτό το κεφάλαιο, παρουσιάζουμε τις βασικές αρχές που εμπλέκονται στην στατιστική ανάλυση χρονοσειρών. Αρχικά, δίνουμε έναν αυστηρό ορισμός της χρονοσειράς. Στην πραγματικότητα, μια χρονοσειρά είναι ένας ειδικός τύπος στοχαστικής διαδικασίας.

## 1.1 Ορισμός

Μια χρονοσειρά είναι μια στοχαστική διαδικασία $\{X(t); t \in T\}$ η οποία αποτελείται από τυχαίες μεταβλητές, όπου $T$ ο χρόνος, για τον οποίο όλες οι τυχαίες μεταβλητές $X(t), t \in T$ ορίζονται στον ίδιο δειγματικό χώρο. Η πραγματοποίηση μιας χρονοσειράς $\{X(t); t \in T\}$ είναι το σύνολο των πραγματικών αποτελεσμάτων $\{X(t, \omega); t \in T\}$ για κάποιο σταθερό $\omega \in \Omega$.

Απλούστερα, η πραγματοποίηση μιας χρονοσειράς είναι το σύνολο των τιμών $\{X(t)\}$ ως αποτέλεσμα της εμφάνισης ενός παρατηρούμενου γεγονότος και συμβολίζεται ως $\{x(t); t \in T\}$.

Αν ο χρόνος αποτελείται από ένα συνεχές εύρος τιμών (π.χ. $T = (0, \infty)$), η χρονοσειρά ονομάζεται συνεχής, ενώ αν ο χρόνος αποτελείταια από διακριτές τιμές (π.χ. $T = \{0, 1, 2, ...\}$) η χρονοσειρά ονομάζεται διακριτή. Παραδείγματα συνεχών χρονοσειρών είναι η συγκέντρωση ενός χημικού συστατικού, οι μετρήσεις πίεσης, η ροή των υδάτων κ.τ.λ.. Αντιθέτως, η παραγωγή ενός εργοστασίου, ο πληθυσμός μιας χώρας, ή οι συναλλαγματικές ισοτιμίες μεταξύ νομισμάτων αποτελούν παραδείγματα διακριτών χρονοσειρών.

Μια χρονοσειρά επηρεάζεται από τέσσερις συνιστώσες, οι οποίες διαχωρίζονται από τα παρατηρούμενα δεδομένα. Αυτές οι συνιστώσες είναι: η συνιστώσα της Τάσης (trend), η Κυκλική συνιστώσα (cyclical), η Εποχική συνιστώσα (Seasonal) και η Τυχαιότητα (random or irregular) (Adhikari R., Agrawal R. (2013)). Πιο αναλυτικά:

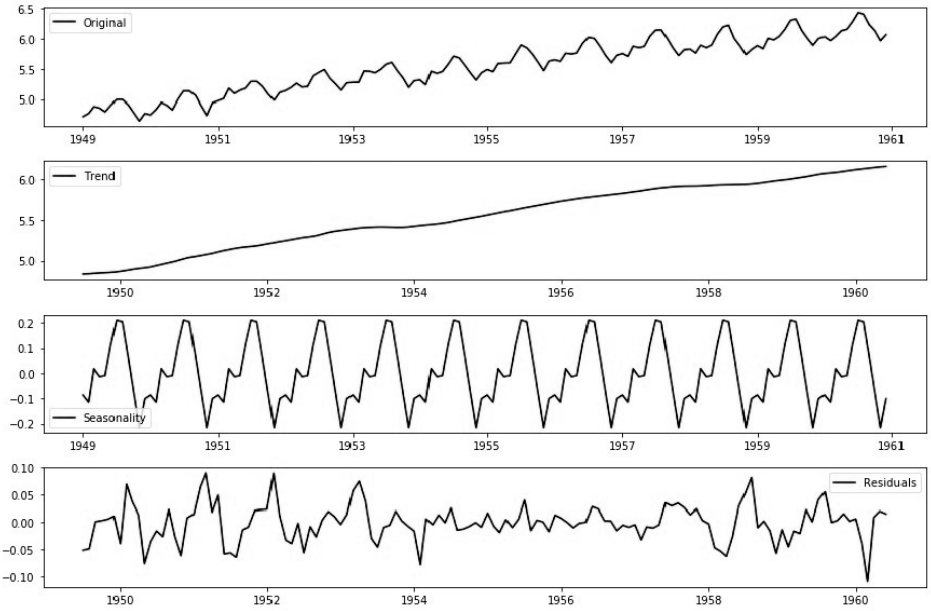
**Τάση:** Αφορά την γενική τάση που έχει μια χρονοσειρά να αυξάνεται, να μειώνεται ή να παραμένει σταθερή με την πάροδο μεγάλου χρονικού διαστήματος.

**Κυκλική συνιστώσα:** Περιγράφει τις μακροπρόθεσμες αλλαγές οι οποίες προκαλούνται από διάφορες καταστάσεις, οι οποίες επαναλαμβάνονται κυκλικά. Η διάρκεια ενός κύκλου έχει μεγάλη διάρκεια (δύο ή περισσότερα χρόνια).

**Εποχική συνιστώσα:** Αφορά τις εποχιακές διακυμάνσεις που παρουσιάζονται σε ένα έτος ή κατά την διάρκεια μιας σεζόν.Ο καιρός, το κλίμα, τα έθιμα, οι παραδόσεις κ.λ.π. αποτελούν κύριους παράγοντες που προκαλόυν εποχικές διακυμάνσεις. Διαφέρει από την κυκλική συνιστώσα, καθώς υπάρχει επανάληψη σε συγκεκριμένο χρονικό διάστημα και όχι επανάληψη σε κύκλους.

**Τυχαία ή ανώμαλη συνιστώσα:** Αφορά απρόβλεπτες επιρροές, ,μη κανονικές, που δεν επαναλαμβάνονται σε ένα συγκεκριμένο μοτίβο (π.χ.πόλεμος, σεισμοί, πλημμύρες).

Με την αφαίερση των τεσσάρων αυτών συνιστωσών από την χρονοσειρά υπολογίζουμε το υπόλοιπο (remainder component).




**Σχήμα 1.1:** Ανάλυση της χρονοσειράς σε συνιστώσες

**•Ορισμοί**

Αν $\{X(t); t \in T\}$ είναι μια χρονοσειρά, τότε για κάθε $t_1, t_2 \in T$, ορίζουμε

1. Την συνάρτηση $\gamma(\cdot)$ (autocovariance function) ως

$$\gamma(t_1, t_2) = E\{[X(t_1) - \mu(t_1)][X(t_2) - \mu(t_2)]\} \tag{1.1}$$

2. Την συνάρτηση αυτοσυσχέτησης $\rho(\cdot)$ ((autocorrelation function) ως

$$\rho(t_1, t_2) = \frac{\gamma(t_1, t_2)}{\sigma(t_1)\sigma(t_2)} \tag{1.2}$$

## 1.2 Στατικές χρονοσειρές

Στη μελέτη μιας χρονοσειράς, είναι σύνηθες να είναι διαθέσιμο μόνο ένα μεμονωμένο $x_t$. Η ανάλυση της χρονολογικής σειράς με βάση μία μόνο πραγματοποίηση είναι ανάλογη με την ανάλυση των ιδιοτήτων μιας τυχαίας μεταβλητής με βάση μια ενιαία παρατήρηση. Η έννοια της στασιμότητας (stationarity) θα παίξει σημαντικό ρόλο στην ανάλυση μια χρονοσειράς

(Woodward W. A., Gray H. L., Elliott A. C. (2017)). Γενικά μια χρονοσειρά είναι στάσιμη (stationary) αν δεν αλλάζει στο χρόνο.

### 1.2.1 Στασιμότητα

Μία διαδικασία $\{X(t); t \in T\}$ καλείται (αυστηρά) στάσιμη αν για κάθε $t_1, t_2, ..., t_k \in T$ και κάθε $h \in T$, η από κοινού κατανομή της $\{X(t_1), X(t_2), ..., X(t_k)\}$ είναι όμοια με αυτή της $\{X(t_1 + h), X(t_2 + h), ..., X(t_k + h)\}$.

Μία χρονοσειρά $\{X(t); t \in T\}$ καλείται στάσιμη αν:

1. $E[X(t)] = \mu$ σταθερή για όλα τα $t$
2. $Var[X(t)] = \sigma^2 < \infty$ πεπερασμένη,σταθερή για όλα τα $t$
3. $\gamma(t_1, t_2)$ εξαρτάται μόνο από το $t_2 - t_1$ και ισχύει $Cov(X(t), X(t + h)) = \gamma(h)$.

Στην συνέχεια αναφέρουμε τις βασικές ιδιότητες της συνάρτησης $\gamma(\cdot)$ και της συνάρτησης αυτοσυσχέτησης $\rho(\cdot)$ για στάσιμες χρονοσειρές.

**Ιδιότητες της συνάρτησης** $\gamma(\cdot)$

1. $\gamma(0) = \sigma^2$
2. $|\gamma(h)| \leq \gamma(0)$ για κάθε h
3. $\gamma(h) = \gamma(-h)$
4. Η συνάρτηση $\gamma(h)$ είναι θετικά ορισμένη.

**Ιδιότητες της συνάρτησης** $\rho(\cdot)$

1. $\rho(0) = 1$
2. $|\rho(h)| \leq$ για κάθε $h$

3. $\rho(h) = \rho(-h)$

4. Η συνάρτηση $\rho(h)$ είναι θετικά ορισμένη και για διακριτές χρονοσειρές ορισμένες για $t = 0, \pm 1, \pm 2, ..,$ ο πίνακας $\rho_k = \begin{pmatrix} 1 & \rho_1 & ... & \rho_k \\ \rho_1 & 1 & ... & \rho_k - 1 \\ \vdots & \vdots & \ddots & \vdots \\ \rho_k & \rho_k - 1 & ... & 1 \end{pmatrix}$ είναι θετικά ορισμένος για κάθε $k$.

Παρακάτω παραθέτουμε τις αποδείξεις των ιδιοτήτων:

1. *Απόδειξη.* $\gamma(0) = Cov(X(t), X(t)) = Var(X(t)) = \sigma^2$.

□

2. *Απόδειξη.* $|\gamma(h)| \leq \gamma(0) \ \forall h$.

   Η ανισότητα μπορεί να αποδειχθεί σημειώνοντας ότι για οποιεσδήποτε τυχαίες μεταβλητές $X, Y$ ισχύει:

   $$E[(X - \mu_X)(Y - \mu_Y)] \leq \{E[(X - \mu_X)^2]E[(Y - \mu_Y)^2]\}^{1/2}$$

   από ανισότητα Cauchy-Schwarz. Τώρα θέτωντας $X = X(t)$ και $Y = X(t+h)$ έχουμε:

   $$|\gamma(h)| \leq E|(X(t) - \mu)(X(t + h) - \mu)| \leq \sigma^2 = \gamma(0) \quad .$$

□

3. *Απόδειξη.* $\gamma(h) = \gamma(-h)$.

   Παρατηρούμε ότι:

   $$\begin{aligned} \gamma(-h) &= E[(X(t) - \mu)(X(t - h) - \mu)] \\ &= E[(X(t - h) - \mu)(X(t) - \mu)] \\ &= E[(X(t_1) - \mu)(X(t_1 + h) - \mu)], \end{aligned}$$

   όπου $t_1 = t - h$. Όμως, αφού η συνάρτηση $\gamma$ δεν εξαρτάται από τον χρόνο $t$, η τελευταία ισότητα ισούται με $\gamma(h)$. □

4. *Απόδειξη.* Μία συνάρτηση $F$ καλείται θετικά ορισμένη αν και μόνο αν:

   $$\sum_{i,j=1}^{n} \alpha_i F(i - j)\alpha_j \geq 0 \forall n, \alpha = (\alpha_1, ..., \alpha_n)^T \in \mathbb{R}^n.$$

Παρατηρούμε πώς:

$$0 \leq Var(\alpha^T X_n) = \alpha^T \Gamma_n \alpha = \sum_{i,j=1}^{n} \alpha_i \gamma_{(i-j)} \alpha_j$$

όπου $X_n = (X_n, ..., X_1)^T$ και

$$\begin{aligned}
\Gamma_n &= Var(X_n) \\
&= \begin{pmatrix} Cov(X_n, X_n) & Cov(X_n, X_n - 1) & ... & Cov(X_n, X_1) \\ Cov(X_{n-1}, X_n) & Cov(X_{n-1}, X_{n-1}) & ... & Cov(X_{n-1}, X_1) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_1, X_n) & Cov(X_1, X_{n-1}) & ... & Cov(X_1, X_1) \end{pmatrix} \\
&= \begin{pmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{n-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{n-1} & \gamma_{n-2} & \cdots & \gamma_0 \end{pmatrix}
\end{aligned}$$

□

Οι αποδείξεις των ιδιοτήτων 2 - 4 της συνάρτησης αυτοσυσχέτισης $\rho(\cdot)$ είναι ανάλογες. Παραθέτουμε την απόδειξη της πρώτης ιδιότητας.

1. *Απόδειξη.* Από ορισμό της συνάρτησης αυτοσυσχέτισης έχουμε:

$$\rho(t_1, t_2) = \frac{\gamma(t_1, t_2)}{\sigma(t_1)\sigma(t_2)}$$

Καθώς η χρονοσειρά είναι στάσιμη ισχύει:

$$\rho(h) = \frac{\gamma(h)}{\sigma^2}$$

Σε συνδυασμό με την ιδιότητα 1 της συνάρτησης $\gamma(\cdot)$ καταλήγουμε στο ζητούμενο:

$$\rho(0) = 1.$$

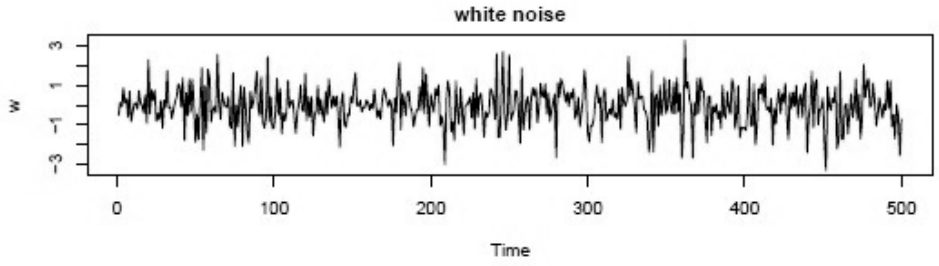
□

## 1.3 Θόρυβος

Το απλούστερο μοντέλο χρονοσειρών είναι αυτό που οι παρατηρήσεις $X_t$ είναι ανεξάρτητες, ισόνομες τυχαίες μεταβλητές με μηδενικό μέσο όρο και δεν παρουσιάζει τάση ή εποχικότητα (Wang D. (2015)). Μια τέτοια ακολουθία $\{X(t)\}$ καλείται Θόρυβος IDD (independent and identically distributed). Συμβολίζεται ως: $X_t \sim IDD(0, \sigma^2)$.

### 1.3.1 Λευκός Θόρυβος

Μια χρονοσειρά καλείται Λευκός Θόρυβος (White Noise) αν ισχύει $Cov(X_{t1}, X_{t2}) = 0$ για κάθε $t_1, t_2$ με $E[X_t] = 0$ και $Var[X_t] = \sigma^2$.
Συμβολίζεται ως $X_t \sim WN(0, \sigma^2)$.
Οι παρατηρήσεις, λοιπόν, είναι ασυσχέτιστες με μέση τιμή 0 και διασπορά ίση με $\sigma^2$.
Κάθε ακολουθία $IDD(0, \sigma^2)$ είναι $WN(0, \sigma^2)$. Το αντίστροφο δεν ισχύει πάντα.




**Σχήμα 1.2:** Χρονοσειρά Λευκός Θόρυβος

# Κεφάλαιο 2

# Πολυμεταβλητός Έλεγχος Ποιότητας

Τα περισσότερα προβλήματα παρακολούθησης και ελέγχου περιλαμβάνουν πολλές συναφείς μεταβλητές. Είναι λοιπόν απαραίτητο να χρησιμοποιούμε πολυμεταβλητές μεθόδους που λαμβάνουν υπόψη από κοινού μεταβλητές. Τα προβλήματα παρακολούθησης διεργασίας, στα οποία αρκετές σχετικές μεταβλητές μας απασχολούν συνήθως καλούνται προβλήματα πολυμεταβλητού ελέγχου ποιότητας (ή παρακολούθησης διεργασίας). Στην συνέχεια της εργασίας παρουσιάζονται διαγράμματα που υπόκεινται στην κατηγορία του Πολυμεταβλητού Ελέγχου Ποιότητας. Σε αυτή την ενότητα εισάγουμε βασικές γνώσεις και εργαλία για τα πολυμεταβλητά διαγράμματα ελέγχου.

## 2.1 Περιγραφή των Πολυμεταβλητών δεδομένων

### 2.1.1 Η πολυμεταβλητή κανονική κατανομή

Υποθέτουμε ότι έχουμε $p$ μεταβλητές, έστω $X_1, X_2, ..., X_p$. Βάζουμε αυτές τις μεταβλητές σε ένα διάνυσμα συνιστωσών $\boldsymbol{X}' = [X_1, X_2, ..., X_p]$. Το διάνυσμα των μέσων των $X$ είναι $\boldsymbol{\mu}' = [\mu_1, \mu_2, ..., \mu_p, ]$. Οι διακυμάνσεις και οι συνδιακυμάνσεις των τυχαίων μεταβλητών περιλαμβάνονται σε έναν πίνακα $p \times p$ που ονομάζεται πίνακας συνδιακύμανσης $\boldsymbol{\Sigma}$ και ορίζεται

ως:

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & ... & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & ... & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & ... & \sigma_{pp} \end{pmatrix}$$

Τα κύρια διαγώνια στοιχεία του πίνακα είναι οι διακυμάνσεις των $\boldsymbol{X}$ και τα στοιχεία εκτός της διαγωνίου είναι οι συνδιακυμάνσεις. Επομένως, η τετραγωνική τυποποιημένη απόσταση του $\boldsymbol{X}$ στο $\boldsymbol{\mu}$ είναι

$$(\boldsymbol{X}-\boldsymbol{\mu})^{'}(\boldsymbol{\Sigma})^{-1}(\boldsymbol{X}-\boldsymbol{\mu})$$

Με βάση τα παραπάνω η συνάρτηση πυκνότητας πιθανότητας της πολυμεταβλητής κανονικής κατανομής δίνεται από την σχέση:

$$f(\boldsymbol{X}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^{'}(\boldsymbol{\Sigma})^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\right) \tag{2.1}$$

με $-\infty < X_j < \infty, j = 1, 2, ..., p$.

Στην περίπτωση που $p = 2$ η πολυμεταβλητή κανονική κατανομή καλείται διμεταβλητή κανονική κατανομή. Στην περίπτωση αυτή, το διάνυσμα του μέσου ισούται με $\mu = \begin{pmatrix} \mu_1 \\ \mu_1 \end{pmatrix}$, και ο πίνακας συνδιακύμανσης $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$.

Τέλος, αξίζει να σημειωθεί ότι η συνάρτηση πυκνότητας πιθανότητας είναι μια επιφάνεια.

### 2.1.2 Διάνυσμα Δειγματικού Μέσου και ο Πίνακας δειγματικής Συνδιακύμανσης

Έστω, ένα τυχαίο δείγμα που προέρχεται από την πολυμεταβλητή κανονική κατανομή

$$\boldsymbol{X_1}, \boldsymbol{X_2}, ..., \boldsymbol{X_n}$$

όπου το $i-$οστό δειγματικό διάνυσμα περιέχει παρατηρήσεις για κάθε μία από τις μεταβλητές. Τότε, το διάνυσμα του δειγματικού μέσου (sample mean vector) είναι:

$$\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X_i} \tag{2.2}$$

και ο δειγματικός πίνακας συνδιακύμανσης (sample covarinace matrix) είναι:

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X_i} - \bar{\boldsymbol{X}})(\boldsymbol{X_i} - \bar{\boldsymbol{X}})'. \tag{2.3}$$

Οι δειγματικές διακυμάνσεις στην κύρια διαγώνιο του πίνακα υπολογίζονται από τη σχέση:

$$S_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2 \tag{2.4}$$

και οι δειγματικές συνδιακυμάνσεις είναι:

$$S_{jk}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)(X_{ij} - \bar{X}_k) \tag{2.5}$$

Το διάνυσμα του δειγματικού μέσου και ο δειγματικός πίνακας συνδιακύμανσης είναι αμερόληπτες εκτιμήτριες του μέσου και της συνδιακύμανσης αφού:

$$E(\bar{\boldsymbol{X}}) = \boldsymbol{\mu} \tag{2.6}$$

$$E(\boldsymbol{S}) = \boldsymbol{\Sigma}. \tag{2.7}$$

## 2.2 Το Hotelling $T^2$ διάγραμμα ελέγχου

Μια από τις βασικότερες διαδικασίες παρακολούθησης είναι το Hotelling $T^2$ διάγραμμα ελέγχου. Είναι ένα πολύ χαρακτηριστικό διάγραμμα, που θα συναντήσουμε και στην συνέχεια. Στην παράγραφο αυτή θα παρουσιάσουμε δύο εκδόσεις του διαγράμματος.

Υποθέτουμε ότι η από κοινού κατανομή πιθανότητας των $p$ ποιοτικών χαρακτηριστικών είναι μια $p$-μεταβλητή κανονική κατανομή. Το σύνολο των μέσων των ποιοτικών

χαρακτηριστικών παριστάνεται με ένα $p \times 1$ διάνυσμα

$$\bar{\boldsymbol{X}} = \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_p \end{pmatrix}$$

Η στατιστική συνάρτηση ελέγχου στο $\chi^2$ διάγραμμα ελέγχου για κάθε δείγμα είναι

$$\chi_0^2 = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}), \tag{2.8}$$

όπου $\boldsymbol{\mu}'$ είναι ο πίνακας των εντός ελέγχου μέσων για κάθε ποιοτικό χαρακτηριστικό και ο πίκανας συνδιακύμανσης $\boldsymbol{\Sigma}$. Το άνω όριο ελέγχου του διαγράμματος ελέγχου είναι:

$$UCL = \chi_{\alpha,p}^2. \tag{2.9}$$

Στην συνέχεια είναι απαραίτητο να εκτιμήσουμε τα $\boldsymbol{\mu}$ και $\boldsymbol{\Sigma}$. Έστω $m$ δείγματα. Οι δειγματικοί μέσοι και οι δειγματικές διακυμάνσεις υπολογίζονται από κάθε δείγμα ως εξής:

$$\bar{X}_{jk} = \frac{1}{n} \sum_{i=1}^{n} X_{ijk}, \quad j = 1, ..., p \quad k = 1, ..., m, \tag{2.10}$$

$$S_{jk}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ijk} - \bar{X}_{jk})^2 \quad j = 1, ..., p \quad k = 1, ..., m. \tag{2.11}$$

Η συνδιακύμανση μεταξύ του ποιοτικού χαρακτηριστικού $j$ και του ποιοτικού χαρτακτηριστικού $h$ στο $k-$οστό δείγμα είναι:

$$S_{jhk}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ijk} - \bar{X}_{hk})^2 \quad k = 1, ..., m \quad j \neq h. \tag{2.12}$$

Μετέπειτα, υπολογίζονται κατά μέσο όρο για όλα τα δείγματα τα στατιστικά $\bar{X}_{jk}$ και $\bar{S}_j^2$ και έχουμε:

$$\bar{\bar{X}}_j = \frac{1}{m} \sum_{k=1}^{m} \bar{X}_{jk}, \quad j = 1, ..., p, \tag{2.13}$$

$$\bar{S}_j^2 = \frac{1}{m}\sum_{k=1}^{m} \bar{S}_{jk}^2, \quad j = 1, ..., p, \tag{2.14}$$

και

$$\bar{S}_{jh} = \frac{1}{m}\sum_{k=1}^{m} \bar{S}_{jhk}, \quad j \neq h. \tag{2.15}$$

Επομένως, σχηματίζουμε το διάνυσμα $\bar{\bar{\boldsymbol{X}}}$ με στοιχεία τα $\{\bar{\bar{X}}_j\}$ και τον μέσο των δειγματικών πινάκων συνσιακύμανσης ως:

$$\boldsymbol{S} = \begin{pmatrix} \bar{S}_1^2 & \bar{S}_{12} & \bar{S}_{13} & ... & \bar{S}_{1p} \\ & \bar{S}_2^2 & \bar{S}_{23} & ... & \bar{S}_{2p} \\ & & \bar{S}_3^2 & ... & \vdots \\ & & & \ddots & \bar{S}_p^2 \end{pmatrix} \tag{2.16}$$

Όταν η διεργασία είναι εντός ελέγχου ο μέσος $\boldsymbol{S}$ είναι αμερόληπτη εκτιμήτρια του $\boldsymbol{\Sigma}$.

Από την σχέση (2.16), χρησιμοποιούμε το $\boldsymbol{S}$ ως εκτιμήτρια του $\boldsymbol{\Sigma}$. Αντικαθιστούμε το $\boldsymbol{\mu}$ με το $\bar{\bar{\boldsymbol{X}}}$ και το $\boldsymbol{\Sigma}$ με το $\boldsymbol{S}$ στην σχέση (2.8) και έχουμε την διαδικασία Hotelling $T^2$ διάγραμμα ελέγχου ως εξής:

$$T^2 = n(\bar{\boldsymbol{X}} - \bar{\bar{\boldsymbol{X}}})' \boldsymbol{S^{-1}} (\bar{\boldsymbol{X}} - \bar{\bar{\boldsymbol{X}}}) \tag{2.17}$$

Είναι σημαντικό να δοθεί μεγάλη προσοχή στην επιλογή των ορίων ελέγχου για το στατιστικό Hotelling $T^2$. Η επιλογή βασίζεται στον τρόπο με τον οποίο χρησιμοποιείται το διάγραμμα. Υπάρχουν δύο διαφορετικές φάσεις της χρήσης του διαγράμματος ελέγχου. Η Φάση I αφορά την εξέταση αν η διεργασία ήταν εντός ελέγχου. Ο στόχος στη Φάση I είναι να λάβουμε ένα εντός ελέγχου σύνολο παρατηρήσεων, έτσι ώστε τα όρια ελέγχου να ορίζονται για τη Φάση II, η οποία είναι η παρακολούθηση της μελλοντικής παραγωγής.

Τα όρια ελέγχου στην Φάση I δίνονται από:

$$\begin{aligned} UCL &= \frac{p(m-1)(n-1)}{mn-m-p+1} F_{\alpha,p,mn-m-p+1} \\ LCL &= 0. \end{aligned} \tag{2.18}$$

Τα όρια ελέγχου στην Φάση ΙΙ δίνονται από:

$$\begin{aligned} UCL &= \frac{p(m+1)(n-1)}{mn-m-p+1} F_{\alpha,p,mn-m-p+1} \\ LCL &= 0. \end{aligned} \tag{2.19}$$

Η δυσκολία που συναντάται εδώ, αλλά και σε οποιοδήποτε πολυμεταβλητό διάγραμμα ελέγχου είναι η στατιστική ερμηνεία ενός εκτός ελέγχου σήματος. Ειδικότερα, ποια ή ποιες από τις μεταβλητές είναι υπεύθυνη/ές για την εκτός ελέγχου ένδειξη· Αυτό το ερώτημα δεν είναι πάντα εύκολο να απαντηθεί. Μια χρήσιμη προσέγγιση για την επίλυση αυτού του προβλήματος στη διάγνωση ενός εκτός ελέγχου σήματος είναι η ανάλυση του στατιστικού $T^2$ σε συνιστώσες που αντανακλούν τη συνεισφορά κάθε μεμονωμένης μεταβλητής. Αν $T^2$ είναι η τρέχουσα τιμή του στατιστικού, και $T^2_{(i)}$ είναι η τιμή του στατιστικού για όλες τις μεταβλητές της διεργασίας δίχως την $i$-οστή, τότε:

$$d_i = T^2 - T^2_{(i)} \tag{2.20}$$

είναι ένας δείκτης της σχετικής συνεισφοράς της $i$-οστή μεταβλητής στο ολικό στατιστικό. Με αυτόν τον τρόπο, όταν παράγεται ένα εκτός ελέγχου σήμα, υπολογίζουμε τα $d_i$. Οι μεταβλητές για τις οποίες η τιμή των $d_i$ είναι σχετικά μεγάλη, πρέπει να εξεταστούν. Αυτή η διαδικασία έχει ένα πρόσθετο πλεονέκτημα, καθώς οι υπολογισμοί μπορούν να πραγματοποιηθούν με χρήση υπολογιστή.

Τέλος, θα εξέτασουμε την περίπτωση του Hotelling $T^2$ με την μέθοδο των μεμονωμένων παρατηρήσεων. Υποθέσουμε ότι έχουμε $m$ δείγματα, μεγέθους $n = 1$ το καθένα και $p$ είναι τα ποιοτικά χαρακτηριστικά που παρατηρούνται σε κάθε δείγμα. Έστω $\bar{\boldsymbol{X}}$ το διάνυσμα του δειγματικού μέσου και $\boldsymbol{S}$ ο πίνακας συνδιακύμανσης των παρατηρήσεων. Το Hotelling $T^2$ στατιστικό της σχέσης (2.17) γίνεται:

$$T^2 = n(\bar{\boldsymbol{X}} - \bar{\boldsymbol{X}})' \boldsymbol{S}^{-1} (\bar{\boldsymbol{X}} - \bar{\boldsymbol{X}}) \tag{2.21}$$

Σε αυτή την περίπτωση τα όρια ελέγχου στην Φάση ΙΙ είναι:

$$\begin{aligned} UCL &= \frac{p(m+1)(m-1)}{m^2-mp} F_{\alpha,p,m-p} \\ LCL &= 0. \end{aligned} \tag{2.22}$$

Όταν ο αριθμός των δειγμάτων $m$ είναι μεγάλος, χρησιμοποιείται το προσεγγιστικό όριο ελέγχου:

$$UCL = \frac{p(m-1)}{m-p} F_{\alpha,p,m-p}, \tag{2.23}$$

Μια ειδική περίπτωση ορίων ισχύει για $n = 1$ όπου τα όρια ελέγχου στην Φάση Ι βασίζονται στην κατανομή Βήτα και έχουν την μορφή:

$$\begin{aligned} UCL &= \frac{(m-1)^2}{m} \beta_{\alpha,p/2,(m-p-1)/2} \\ LCL &= 0. \end{aligned} \tag{2.24}$$

όπου $\beta_{\alpha,p/2,(m-p-1)/2}$ το άνω $\alpha$ ποσοστημόριο της κατανομής Βήτα με παραμέτρους $p/2, (m-p-1)/2$.

# Κεφάλαιο 3

# Τεχνικές Μηχανικής Μάθησης στα Διαγράμματα Ελέγχου για Συσχετισμένα Δεδομένα

Σε αυτό το κεφάλαιο εισάγουμε τις τεχνικές Μηχανικής Μάθησης (Machine Learning - ML) στον στατιστικό έλεγχο διεργασιών (ΣΕΔ) και στα διαγράμματα ελέγχου. Η χρήση διαγραμμάτων ελέγχου που βασίζονται στην Μηχανική Μάθηση κρίνεται αναγκαία καθώς οι υποθέσεις ανεξαρτησίας των δεδομένων, στην πλέον πολυσύνθετη διαδικασία παραγωγής, είναι εσφαλμένες και οδηγούν σε ψευδείς πληροφορίες. Αφού κάνουμε αναφορά σε βασικές τεχνικές Μηχανικής Μάθησης, θα περιγράψουμε αντιπροσωπευτικά διαγράμματα ελέγχου που βασίζονται σε αυτές και τέλος θα αναλύσουμε μια διαδικασία αποσυσχέτισης των δεδομένων και πως αυτή εφαρμόζεται στα διαγράμματα που θα έχουμε αναφέρει.

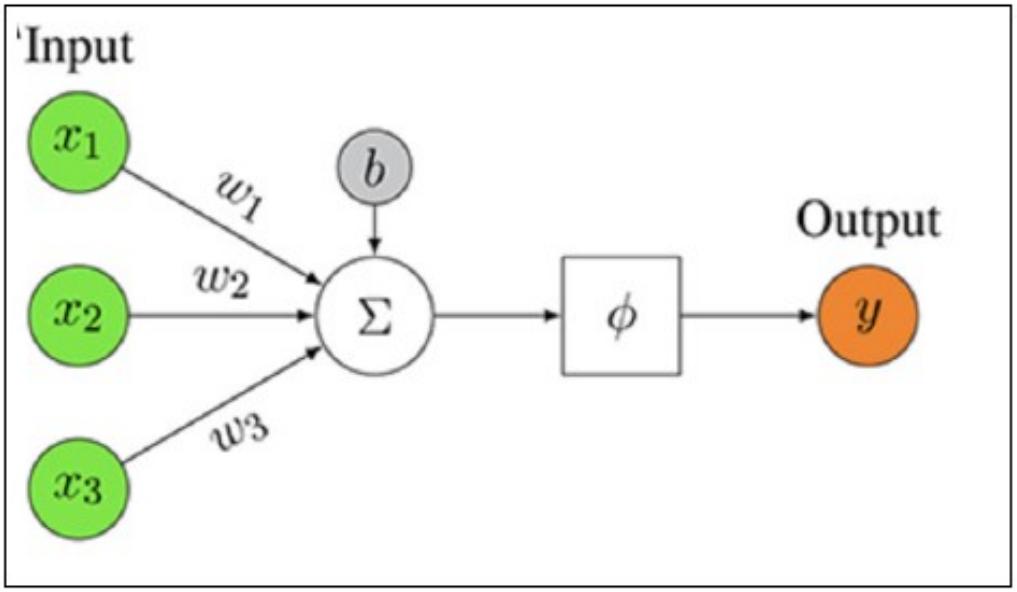
## 3.1 Βασικές Τεχνικές Μηχανικής Μάθησης

Η Μηχανική Μάθηση είναι ένας τομέας της Τεχνητής Νοημοσύνης (Artificial Intellingence - AI), ο οποίος αποτελείται από προγραμματιστικούς αλγορίθμους που μαθαίνουν αυτόματα από δεδομένα και εμπειρίες ή μέσω αλληλεπίδρασης με το περιβάλλον (Viharos Z. J., Jakab R. (2021)). Αυτό που κάνει την ML πραγματικά χρήσιμη είναι το εξής γεγονός: ο αλγόριθ-

μος μπορεί να «μάθει» και να προσαρμόσει τα αποτελέσματά του με βάση νέα δεδομένα χωρίς εκ των προτέρων προγραμματισμό. Υπάρχουν τρείς κύριοι κλάδοι: Επιβλεπόμενης Μάθηση (supervised learning), μη Επιβλεπόμενη Μάθηση (unsupervised learning) και Ενισχυτική Μάθηση (reinforcement learning). Πιο αναλυτικά, ο αλγόριθμος της Επιβλεπόμενης Μάθησης χρησιμοποιείται για την εύρεση συσχετισμών μεταξύ των εισαγώμενων δεδομένων (επεξηγηματικές μεταβλητές) και των δεδομένων εξόδου (προβλέψιμες μεταβλητές), έτσι ώστε να μπορούμε να προβλέψουμε τις τιμές εξόδου για νέα δεδομένα με βάση τις σχέσεις που έμαθε ο αλγόριθμος από τα προηγούμενα σύνολα δεδομένων. Αντιθέτως, η τεχνική της μη Επιβλεπόμενης Μάθησης πρέπει μόνη να ανακαλύψει κάποια πιθανή δομή που μπορεί να κρύβεται πίσω από μη χαρακτηρισμένα δεδομένα. Τέλος, η Ενισχυτική Μάθηση είναι ένας τομέας της ML που ασχολείται με τον τρόπο λήψης μιας σειράς αποφάσεων. Τέτοιες τεχνικές συνδιάζονται όλο και περισσότερο με τα διαγράμματα ελέγχου. Στην συνέχεια, αναφέρουμε κάποιες βασικές τεχνικές που χρησιμοποιούνται στον σχεδιασμό διαγραμμάτων ελέγχου.

### 3.1.1 Νευρωνικά Δίκτυα και Βαθιά Μάθηση

Τα Νευρωνικά Δίκτυα (Neural Networks - NN) αποτελούν την πιο αποτελεσματική μέθοδο εκμάθησης, χάρη στην ικανότητά τους να μοντελοποιούν ένα σύστημα σε πραγματικό χρόνο. Αποτελούν ένα υπολογιστικό μοντέλο που προσομοιώνει τα νευρικά κύτταρα του ανθρώπου. Τα NN κατασκευάζονται από στρώματα συνδεδεμένων μονάδων που ονομάζονται νευρώνες. Συγκεκριμένα υπάρχουν το στρώμα εισόδου, το κρυφό στρώμα και το στρώμα εξόδου. Η βασική ιδέα για την λειτουργία ενός νευρώνα είναι η εξής: Τα δεδομένα εισόδου, έστω $x$, με μια μεροληψία $b$ σταθμίζονται με βάρη $w$ και ύστερα αθροίζονται. Τα $x$ και $w$ είναι διανύσματα με $x, w \in \mathbb{R}^n$ όπου $n \in \mathbb{N}$ είναι η διάσταση της εισόδου. Η μεροληψία είναι βαθμωτό μέγεθος. Το άθροισμα αυτών των όρων, έστω $z$, ορίζεται ως: $z = w^T x + b$. Το $z$ εισέρχεται σε μια συνάρτηση ενεργοποίησης $\phi$ και το αποτέλεσμα είναι είτε το δεδομένο εισόδου του επόμενου στρώματος, είτε το δεδομένο εξόδου. Για την εκπαίδευση του NN, το σύνολο δεδομένων χωρίζεται σε δεδομένα εκπαίδευσης και δοκιμής. Το NN πρέπει να εκπαιδευτεί με ορισμένες τεχνικές εκμάθησης, όπως η οπίσθια διάδοση Levenberg–Marquardt (trainlm) για να επιτευχθεί το καλύτερο αποτέλεσμα. Το νευρωνικό δίκτυο αναδρομι-




**Σχήμα 3.1:** Λειτουργεία ενός νευρώνα

κής διάδοσης (Back Propagation Neural Network - BPNN) είναι ο πιο δημοφιλής τύπος νευρωνικού δικτύου. Το BPNN χρησιμοποιεί μια επιβλεπόμενη μέθοδο μάθησης. Τα δεδομένα εκπαίδευσης επιλέγονται τυχαία από συνδυασμούς εισόδων και εξόδων. Ένα καλά εκπαιδευμένο μοντέλο NN έχει την ικανότητα να ορίζει μια σχέση μεταξύ εισόδων και εξόδων χωρίς να υπάρχει μαθηματική σχέση. Εάν το σφάλμα φτάσει στην ελάχιστη τιμή, η διαδικασία διακόπτεται. Διαφορετικά, τροποποιούνται τα βάρη σύνδεσης για καλύτερα αποτελέσματα. Μεταξύ των μοντέλων NN, τα αναδρομικά νευρωνικά δίκτυα (Recurrent Neural Network - RNN) είναι ιδιαίτερα χρήσιμα στη μοντελοποίηση χρονοσειρών καθώς επιτρέπουν την σύνδεση νευρώνων που βρίσκονται στο ίδιο κρυφό στρώμα. Αποτελούνται απο τρία κύρια στοιχεία, τα δεδομένα εισόδου $x$, το κρυφό στρώμα $h$ και τα δεδομένα εξόδου $o$. Το $(x_{t,1}, x_{t,2}, ..., x_{t,n})$ με $n$ δεδομένα, είναι η ακολουθία εισόδου του RNN για χρόνο $t$. Επίσης, $W^{(hx)}$ είναι ο πίνακας των βάρεων του κρυφού στρώματος με το στρώμα εισόδου και $W^{oh}$ ο πίνακας των βάρεων του κρυφού στρώματος με του στρώματος εξόδου. Σε αντίθεση με άλλα NN, υπάρχουν επαναλαμβανόμενες συνδέσεις ανάμεσα στο κρυφό στρώμα και τον εαυτό του σε διπλανά χρονικά βήματα στο RNN. Σε χρόνο $t$, αυτοί οι κόμβοι στο κρυφό στρώμα $h_t$ λαμβάνουν την είσοδο των τρεχουσών χρονοσειρών $(x_{t,1}, x_{t,2}, ..., x_{t,n})$ και την τιμή των κόμβων του κρυφού στρώματος στην προηγούμενη κατάσταση $h_{t-1}$. Η κρυφή τιμή κόμβου $h_t$ τροφοδοτείται στη συνέχεια στο στρώμα εξόδου για να δημιουργήσει την τιμή

εξόδου κάθε χρόνο $t$ $o_t$. Έτσι, η είσοδος σε χρόνος $t1$ $(x_{t-1,1}, x_{t-1,2}, ..., x_{t-1,n})$ μπορεί να επηρεάσει την έξοδο τη στιγμή $t$ $o_t$ λόγω των επαναλαμβανόμενων συνδέσεων. Το $h_t$ και το $o_t$ υπολογίζονται ως εξής:

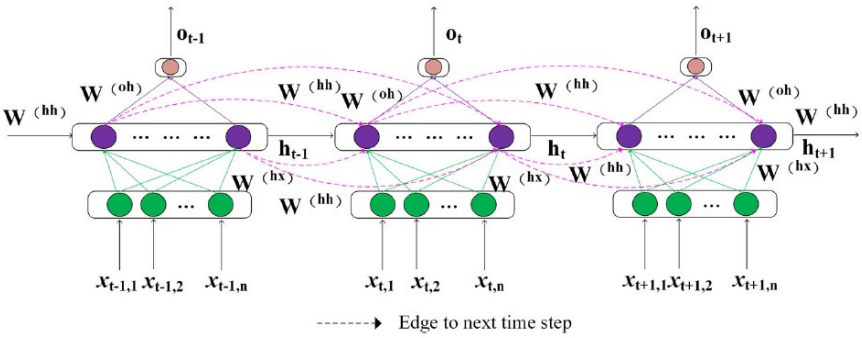
$$h_t = f\left(W^{hh}h_{t-1} + W^{hx}x_t + b_h\right) \tag{3.1}$$

$$o_t = f\left(W^{oh}h_t + b_o\right) \tag{3.2}$$

Για να δούμε κατά πόσο οι προβλεπόμενες τιμές απέχουν από τις πραγματικές χρησιμοποιούμε μια συνάρτηση κόστους. Στο RNN χρησιμοποιείται το σφάλμα διασταυρούμενης εντροπίας (cross entropy error) σε μια ακολουθία μεγέθους $T$, η οποία υπολογίζεται ως:

$$J = -\frac{1}{T}\sum_{t=1}^{T} J^{(t)}(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{|V|} o_{i,j} x log(\hat{o}_{t,j}) \tag{3.3}$$

όπου $|V|$ το μέγεθος του στρώματος εξόδου, $\hat{o}_{t,j}$ είναι η προβλεπόμενη έξοδος του $j$-οστού νευρώνα του στρώματος εξόδου και το $o_{i,j}$ είναι η αναμενόμενη έξοδος του $j$-οστού νευρώνα του στρώματος εξόδου.




**Σχήμα 3.2:** Η βασική αρχιτεκτονική του RNN [wileyonlinelibrary.com]

Στα NN, καθώς ο αριθμός των στρωμάτων αυξάνεται, το δίκτυο γίνεται περιπλοκότερο. Εδώ εισάγεται και η Βαθιά Μάθηση (Deep Learning - DL) η οποία περιλαμβάνει πολλαπλά κρυφά στρώματα νευρωνικών δικτύων. Λαμβάνει υπόψη τη μη γραμμική επεξεργασία σε πολλά επίπεδα και τους ελεγχόμενους ή μη ελεγχόμενους παράγοντες μάθησης. Η τεχνική λειτουργεί λαμβάνοντας την έξοδο του προηγούμενου στρώματος ως είσοδο. Έχει αποδειχθεί ότι είναι επιτυχής στην επίλυση πολύπλοκων δομών και εφαρμόζεται ευρέως.

### 3.1.2 Τυχαία Δάση (Random forests)

Ένα τυχαίο δάσος (random forest - RF) είναι ένας ταξινομητής που αποτελείται από μια συλλογή ταξινομητών με δομή δέντρου $\{C(x, \Theta_i), i = 1, 2, ...\}$ όπου τα $\{\Theta_i\}$ είναι ανεξάρτητα και ομοίως κατανεμημένα τυχαία διανύσματα. Κάθε δέντρο δίνει μια ψήφο στην πιο δημοφιλή κλάση στην είσοδο $x$. Το RF, που χρησιμοποιεί μια τυχαία επιλογή χαρακτηριστικών, συνδιάζει την χρήση δύο μεθόδων, της τυχαίας επιλογής εισόδου (random input selection) και της μεθόδου bagging, στην οποία οφείλει και την καινοτομία του. Η γενική ιδέα του bagging είναι η εξής: Αφού δημιουργούμε bootstrap αντίγραφα από το σύνολο εκπαίδευσης, κάθε ταξινομητής εκπαιδεύεται στο συγκεκριμένο σύνολο. Στην συνέχεια, κάθε ταξινομητής προβλέπει μια μεταβλητή εξόδου για κάθε διάνυσμα εισόδου και το αποτέλεσμα λαμβάνεται με πλειοψηφική απόφαση, δηλαδή το τιμή με τις περισσότερες ψήφους επιλέγεται ως η μεταβλητή απόκρισης. Επομένως, τα σύνολα εκπαίδευσης των RF είναι δείγματα bootstrap. Στην συνέχεια δημιουργείται νέο δέντρο για κάθε ένα από τα σετ δεδομένων εκπαίδευσης χρησιμοποιώντας τυχαία επιλογή εισόδου. Δηλαδή, σε κάθε κόμβο ένα μικρό υποσύνολο χαρακτηριστικών επιλέγεται τυχαία για διαχωρισμό. Αυτή η διαδικασία επαναλαμβάνεται μέχρις το δέντρο να φτάσει το μέγιστος μέγεθος. Ο αριθμός των μεταβλητών, έστω $F$, πρέπει να καθοριστεί προηγουμένως. Το σφάλμα του τυχαίου δάσους εξαρτάται από την ποικιλομορφία και την ακρίβεια των μεμονωμένων δέντρων. Όσο υψηλότερη είναι η τιμή του $F$, τόσο μεγαλύτερη είναι η δύναμη ή η ακρίβεια, αλλά τόσο μικρότερη είναι η ποικιλομορφία μεταξύ των μεμονωμένων δέντρων. Από την άλλη πλευρά, όσο χαμηλότερη είναι η τιμή του $F$, τόσο μικρότερη είναι η συσχέτιση μεταξύ των επιμέρους δέντρων. Επομένως, η παράμετρος $F$ είναι η πιο κρίσιμη σε ένα τυχαίο δάσος.
Ο αλγόριθμος για την κατασκευή τυχαίων δασών μπορεί να συνοψιστεί ως εξής:

1. Ορίζουμε τον αριθμό των δέντρων που θα αναπτυχθούν.

2. Για κάθε δέντρο:

   α) Σχεδιάζουμε ένα τυχαίο υποσύνολο ($T_k$) του συνόλου εκπαίδευσης $T$ ($N$ παρατηρήσεις με αντικατάσταση) για να εκπαιδεύσουμε κάθε δέντρο. Τα στοιχεία στο $T$ που δεν ανήκουν στο $T_k$ ονομάζονται out-of-bag (oob).

   β) Ορίζουμε τον $F$ (αριθμός μεταβλητών για διαχωρισμό) $<<$ $p$ (αριθμός μεταβλητών εισόδου) και επιλέγουμε τον καλύτερο διαχωρισμό μεταξύ των τυχαία επιλεγμένων

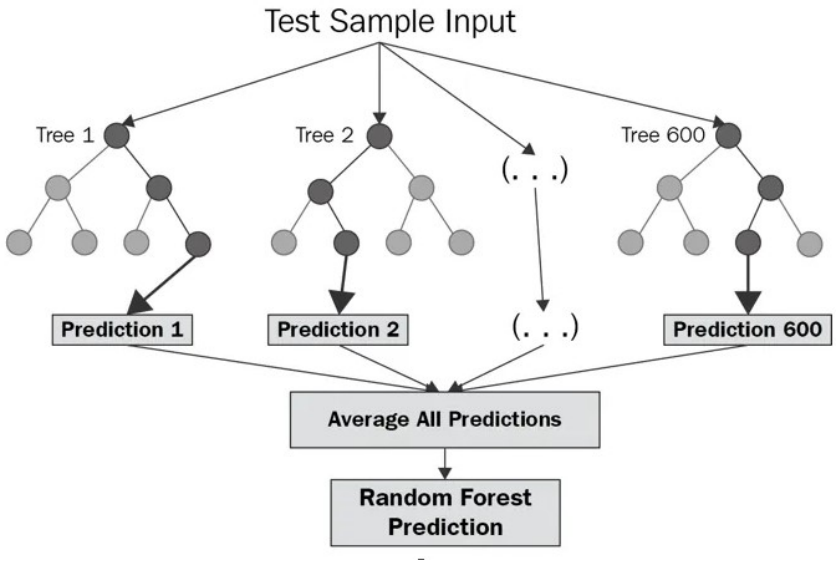
μεταβλητών $F$ για κάθε κόμβο σε κάθε δέντρο.

3. Μεγαλώνουμε το δέντρο στο μέγιστο μέγεθος.

4. Χρησιμοποιούμε τα δεδομένα εκπαίδευσης obb για να εκτιμήσουμε το σφάλμα και τη σημαντικότητα της μεταβλητής.

5. Αναθέτουμε μια κλάση στα νέα δεδομένα, όπως ψήθισε η πλειοψηφία των δέντρων.

6. Χρησιμοποιούμε τα obb δεδομένα για να υπολογίσουμε την ακρίβεια της ταξινόμησης (ή το σφάλμα) για το τυχαίο δάσος και το μέτρο σημασίας για κάθε μεταβλητή εισόδου.




**Σχήμα 3.3:** Δομή RF [corporatefinanceinstitute.com]

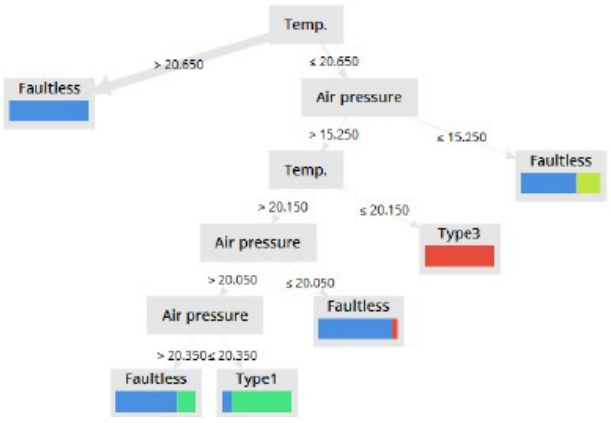
## 3.2 Δέντρα Αποφάσεων (Decision Trees)

Τα δέντρα αποφάσεων (Decision Trees - DT) είναι ένας ευρέως γνωστός ταξινομητής μεταξύ των τεχνικών μηχανικής μάθησης. Είναι σε θέση να χωρίσουν το χώρο σε ορθογώνια, δηλαδή παράγουν όρια που είναι παράλληλα με τους άξονες. Στόχος τους είναι να προβλέψουν τη μεταβλητή στόχο με βάση διάφορες μεταβλητές εισόδου. Υπάρχουν τρεις τύποι κόμβων σε ένα δέντρο: η ρίζα, ο μη τερματικός κόμβος και το φύλλο. Το δέντρο απόφασης

ξεκινά με τον κόμβο της ρίζας, ο οποίος καθορίζεται από το κριτήριο της εντροπίας. Ο μη τερματικός κόμβος και τα φύλλα διαχωρίζονται για να ξεκινήσουν με την υψηλότερη τιμή εντροπίας. Ο τύπος εντροπίας είναι ο εξής:

$$H(X) = -\sum_{i=1}^{n} p(x_i) log_2 p(x_i) \tag{3.4}$$

όπου $X$ μια διακριτή τυχαία μεταβλητή που παίρνει $n$ θετικές τιμές $(x_1, ..., x_n)$.
Η ρίζα, η μη τερματική ρίζα και το φύλλο αντιπροσωπεύουν ένα υποσύνολο των περιπτώσεων εκπαίδευσης. Οι δύο πρώτες ρίζες και οι δύο πρώτοι μη τερματικοί κόμβοι περιλαμβάνουν μία δοκιμή για την τιμή της μεταβλητής. Οι περιπτώσεις εκπαίδευσης χωρίζονται σε δύο ή περισσότερα υποσύνολα ανάλογα με τα αποτελέσματα της δοκιμής. Το δέντρο, στην συνέχεια, κλαδεύεται με αφαίρεση κλαδιών με μικρή στατιστική εγκυρότητα. Ένα δέντρο αποφάσεων είναι βασικά ένα διάγραμμα που μοιάζει με διάγραμμα ροής, με γραμμές που συνδέουν τα σημεία όπου λαμβάνεται μια απόφαση. Οι κανόνες ταξινόμησης είναι κάθε διαδρομή να οδηγεί από κόμβο ρίζας σε κόμβο φύλλου. Η εφαρμογή και η ερμηνεία τους είναι εύκολη, επομένως αποτελούν μια καλή επιλογή για πολλά προβλήματα. Επιπλέον, το δέντρο είναι κατανοητό με τους κανόνες If-Then, επομένως μπορεί να προτιμηθεί έναντι δυσκολότερων τεχνικών. Τα DT μπορεί να παράγουν καλά αποτελέσματα στο σετ εκπαίδευσης, αλλά δεν είναι τόσο καλά στην πρόβλεψη, λόγω overfitting. Η λύση σε αυτό το πρόβλημα είναι να συγκεντρωθούν πολλά δέντρα ώστε για να παραχθούν καλά αποτελέσματα και σε ένα σετ εκπαίδευσης. Αυτό επιτυγχάνεται με την μέθοδο των RF.




**Σχήμα 3.4:** Παράδειγμα Δέντρου Απόφασης

## 3.3 Διαγράμματα Ελέγχου με Τεχνικές Μηχανικής Μάθησης

Έχοντας επισημάνει τις βασικές Τεχνικές Μηχανικής Μάθησης (ML), είμαστε σε θέση να περιγράψουμε τα διαγράμματα ελέγχου τα οποία βασίζονται σε αυτές. Όπως έχουμε αναφέρει, η ταξινόμηση είναι ένας από τους κύριους σκοπούς της επιβλεπόμενης ML και πολλοί αλγόριθμοι όπως οι NN και RF έχουν παρουσιάσει καλή απόδοση στην ακριβή ταξινόμηση δεδομένων εισόδου μετά την εκμάθηση της δομής των δεδομένων από πολλά δεδομένα εκπαίδευσης. Τα ΣΕΔ προβλήματα μπορούν να καταταχθούν σε πρόβλημα ταξινόμησης δυαδικών (binary) κλάσεων. Σε κάθε διαδικασία τα δεδομένα ταξινομούνται είτε εντός ελέγχου (in-control - IC) είτε εκτός ελέγχου (out-of-control - OOC). Πολλοί ML αλγόριθμοι χρειάζονται και τα δύο είδη ιστορικών δεδομένων. Πολλές φορές όμως δεν είναι δυνατόν να έχουμε τις εκτός ελέγχου παρατηρήσεις διαθέσιμες. Μια παραγωγική διαδικασία προσαρμόζεται σωστά κατά την διάρκεια της φάσης εκπαίδευσης (Φάση I) και ένα σύνολο από IC δεδομένων συλλέγονται στην συνέχεια για την εκτίμηση της κατανομής της εντός ελέγχου διαδικασίας ή ορισμένων παραμέτρων της. Επομένως, για τέτοιες εφαρμογές, τα IC δεδομένα είναι διαθέσιμα πριν την Φάση II , αλλά οι OOC παρατηρήσεις είναι συνήθως μη διαθέσιμες. Για να ξεπεραστεί αυτό το πρόβλημα, έχουν προταθεί ιδέες όπως η τεχνητή αντίθεση (artificial contrast), η αντίθεση πραγματικού χρόνου (real-time contrast) και η ταξινόμηση μίας τάξης (one class classification) για την ανάπτυξη διαγραμμάτων ελέγχου χωρίς την υπόθεση της διαθεσιμότητα των OC παρατηρήσεων κατά το στάδιο σχεδιασμού των σχετικών διαγραμμάτων.

### 3.3.1 Διάγραμμα Ελέγχου με Τεχνητή Αντίθεση

Για να ξεπεραστεί το αδιέξοδο της έλλειψης των εκτός ελέγχου παρατηρήσεων, προτάθηκε η χρήση της Τεχνητής Αντίθεσης. Σε αυτή την προσέγγιση, δημιουργείται ένα τεχνητό σύνολο δεδομένων από μία εκτός στόχου κατανομή (π.χ. ομοιόμορφη) και οι παρατηρήσεις σε αυτό το σύνολο δεδομένων θεωρούνται οι εκτός ελέγχου παρατηρήσεις. Μετά, ένας ML αλγόριθμος, π.χ. RF, εφαρμόζεται στα δεδομένα εκπαίδευσης τα οποία αποτελούνται:

1. από τα αυθεντικά εντός ελέγχου δεδομένα, έστω $X_{IC}$ και
2. από το τεχνητό σύνολο δεδομένων, έστω $X_{AC}$.

Στην συνέχεια, η ταξινόμηση πραγματοποιείται από τον αλγόριθμο RF για παρακολούθηση της διαδικασίας.

Τέτοια διαγράμματα ελέγχου, που χρησιμοποιούν RF αλγορίθμους έχουν 2 μεγάλους περιορισμούς:

- Τα ποσοστά σφάλματος ταξινόμησης τους δεν μπορούν να μεταφερθούν στο παραδοσιακό μέσο μήκος ροής ARL (average run length) χωρίς την υπόθεση ανεξαρτησίας των δεδομένων.
- Οι αποφάσεις τους σε μία δεδομένη χρονική στιγμή παίρνονται με βάση τα δεδομένα που παρατηρήθηκαν εκείνη τη στιγμή και μόνο, χωρίς να γίνεται χρήση ιστορικών δεδομένων.

Για να ξεπεραστούν αυτά τα προβλήματα οι Hu και Runger (2010) προτείνουν την εξής μετατροπή που αποτελείται από δύο βήματα:

i Για παρατηρήσεις $\mathbf{X}_n$ για δοσμένη χρονική στιγμή $n$, ο λόγος λογαριθμικής πιθανοφάνειας υπολογίζεται από:

$$l_n = log[\hat{p}_1(\mathbf{X}_n)] - log[\hat{p}_0(\mathbf{X}_n)], \tag{3.5}$$

όπου $\hat{p}_1, \hat{p}_0$ είναι οι εκτιμώμενες πιθανότητες του $\mathbf{X}_n$ σε κάθε τάξη, που λαμβάνεται από τον RF ταξινομητή.

ii Στην συνέχεια προτείνεται ένα τροποποιημένο EWMA διάγραμμα με το εξής στατιστικό:

$$E_n = \lambda l_n + (1 - \lambda) E_{n-1}, \tag{3.6}$$

όπου $\lambda \in (0, 1]$ είναι η παράμετρος στάθμισης βαρύτητας.

Αυτό το διάγραμμα ελέγχου είναι γνωστό σαν AC (artificial contrast). Το όριο ελέγχου του AC διαγράμματος μπορεί να καθοριστεί από την εξής διαδικασία: Αρχικά, το 90% των δεδομένων $X_{IC}$ και τα δεδομένα της τεχνητής αντίθεσης $X_{AC}$ χρησιμοποιούνται για να εκπαιδεύσουν τον ταξινομητή RF. Μετά, το $E_n$ με ένα όριο ελέγχου, έστω $h$ εφαρμόζεται στο υπόλοιπο 10% του συνόλου των $X_{IC}$ για να πάρουμε μια τιμή μήκους ροής (run length - RL). Η παραπάνω διαδικασία επαναλαμβάνεται για 1000 φορές και το μέσο των RL τιμών χρησιμοποιείται για να προσεγγίσουμε το εντός ελέγχου ARL ($ARL_0$) για δοσμένο $h$. Τέλος, το $h$ μπορεί να βρεθεί από έναν αριθμητικό αλγόριθμο.

### 3.3.2 Διάγραμμα Ελέγχου με Αντίθεση Πραγματικού Χρόνου

Τα $X_{AC}$ που χρησιμοποιούνται στο AC διάγραμμα, παράγονται από μία υποκειμενικά εκτός στόχου κατανομή και έτσι μπορεί να μην αντιπροσωπεύει τις πραγματικές εκτός ελέγχου παρατηρήσεις. Κατ' επέκταση ο ταξινομητής που εκπαιδεύεται χρησιμοποιώντας τα $X_{IC}$ και τα $X_{AC}$ μπορεί να μην είναι ο σωστός για την παρακολούθηση μιας συγκεκριμένης διαδικασίας. Για την βελτίωση του AC διαγράμματος προτείνεται μια αντίθεση πραγματικού χρόνου (a real time contrast - RTC). Σε αυτή την προσέγγιση οι πιο πρόσφατες παρατηρήσεις μέσα σε ένα κινούμενο παράθυρο (moving window) του τρέχοντος χρονικού σημείου, χρησιμοποιούνται ως αντιθέσεις. Το σύνολο των εντός ελέγχου δεδομένων χωρίζεται σε δύο μέρη: μία τυχαία επιλεγμένη $N_0$ παρατήρηση από τα $X_{IC}$, έστω $X_{IC_0}$, και χρησιμοποιείται για την εκπαίδευση του RF ταξινομητή. Τα υπόλοιπα δεδομένα, έστω $X_{IC_1}$, χρησιμοποιούνται για τον προσδιορισμό του ορίου ελέγχου. Οι παρατηρήσεις της διεργασίας σε ένα παράθυρο του τρέχοντος χρονικού σημείου $n$ θεωρούνται σαν τα εκτός ελέγχου δεδομένα, έστω $X_{AC_n} = \{\mathbf{X}_{n-w+1}, \mathbf{X}_{n-w+2}, ..., \mathbf{X}_n\}$, όπου $w$ είναι το μέγεθος του παραθύρου. Μετά, ο ταξινομητής RF μπορεί να εκπαιδευτεί ξανά, διαδοχικά με την πάροδο του χρόνου χρησιμοποιώντας το σύνολο δεδομένων που συνδιάζει τα $X_{IC_0}$ και $X_{AC_n}$.

Τα ζεύγη δειγμάτων που δεν περιέχουν κάποια παρατήρηση $\mathbf{X}$ ονομάζονται δείγματα 'out-of-bag' (oob) του $\mathbf{X}$. Για την εκτίμηση του 'out-of-bag' δείκτη ταξινόμησης για τις παρατηρήσεις $X_{IC_0}$ χρησιμοποιούμε το στατιστικό:

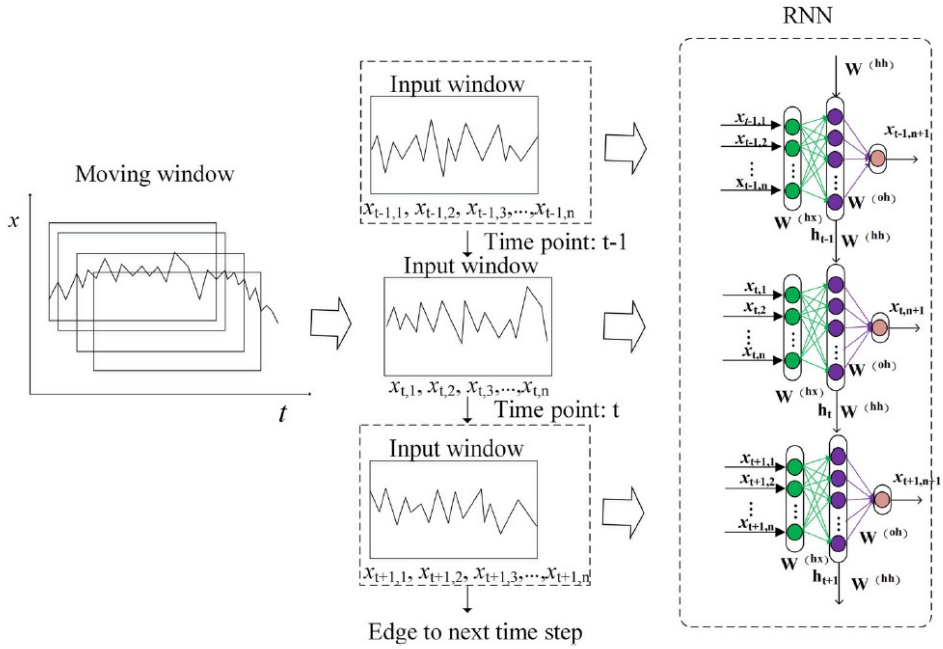
$$P_n = \frac{\sum P_{OOB}(\mathbf{X}_i) I(\mathbf{X_i} \in X_{IC_0})}{|X_{IC_0}|}, \tag{3.7}$$

όπου $|X_{IC_0}|$ ο αριθμός των παρατηρήσεων στο σύνολο $X_{IC_0}$ και $P_{OOB}(\mathbf{X}_i)$ η εκτίμηση της 'out-of-bag' πιθανότητας ταξινόμησης για τα εντός ελέγχου δεδομένα που παίρνουμε από τον ταξινομητή RF.

Τα όρια ελέγχου του παραπάνω διαγράμματος, που πήρε το όνομα RTC (Real Time Contrast), καθορίζονται από μια bootstrap διαδικασία. Αρχικά, επιλέγουμε με αντικατάσταση ένα δείγμα από το σύνολο δεδομένων $X_{IC_1}$. Στην συνέχεια, το RTC διάγραμμα με όριο ελέγχου,έστω $h$, εφαρμόζεται στο δείγμα για να ληφθεί μία τιμή RL. Αυτή η διαδικασία επαναλαμβάνεται για 1000 φορές. Το μέσο RL χρησιμοποιείται για να βρούμε την κατά προσέγγιση τιμή του $ARL_0$ για δεδομένο $h$. Το $h$ επιλέγεται εμπειρικά έτσι ώστε να πραγματοποιηθεί η τιμή του $ARL_0$ και αναζητείται με αριθμητικό αλγόριθμο.

### 3.3.3 Διάγραμμα Υπολοίπων με RNN

Με την χρήση αναδρομικών νευρωνικών δικτύων (RNN) δημιουργούμε ένα διάγραμμα υπολοίπων (residual chart) για τον εντοπισμό των μετατοπίσεων του μέσου των αυτοσυσχετιζόμενων διεργασιών. Αρχικά, πρέπει να κατασκευαστεί ένα RNN κατάλληλο για το διάγραμμα. Η δομή του δικτύου και οι παράμετροι εκμάθησης, προκειμένου να αποκτήσουμε ένα καλά εκπαιδευμένο RNN, ορίζονται χρησιμοποιώντας τη μέθοδο δοκιμής και σφάλματος (trail-error), μια μέθοδος επίλυσης προβλημάτων στην οποία γίνονται πολλαπλές προσπάθειες έως ότου βρεθεί μια λύση. Ακόμη, η μέθοδος κινούμενου παραθύρου (moving window method) χρησιμοποιείται για την παραγωγή των δεδομένων. Περιλαμβάνει $n$ σημεία δεδομένων που συλλέγονται από χρονοσειρές. Ο αριθμός των νευρώνων στο στρώμα εισόδου εξαρτάται από το μέγεθος του παραθύρου $n$. Το μέγεθος του παραθύρου έχει σημαντική επίδραση στην απόδοση του μοντέλου. Ένα μικρό μέγεθος παραθύρου είναι συνήθως ικανό να εντοπίζει ανώμαλα σήματα με μεγάλο μέγεθος μετατόπισης γρήγορα, γεγονός που οδηγεί σε μικρή τιμή του ARL, δηλαδή σε μεγάλο σφάλμα τύπου I. Ένα μεγάλο παράθυρο περιέχει περισσότερα σημεία για ανίχνευση μη φυσιολογικών σημάτων με μικρό μέγεθος μετατόπισης αλλά μπορεί να αυξήσει τον απαιτούμενο χρόνο εντοπισμού τους και να έχουμε μεγάλο ARL και μεγάλο σφάλμα τύπου II. Ως εκ τούτου, είναι απαραίτητο να βρεθεί ένα μέγεθος παραθύρου για την εξισορρόπηση του σφάλματος Τύπου I και Τύπου II. Στην περίπτωση που αναφερόμαστε εμείς, το μέγεθος του παραθύρου υπολογίζεται με την μέθοδο




**Σχήμα 3.5:** Η διαδικασία κατασκευής RNN [wileyonlinelibrary.com]

δοκιμής και σφάλματος. Συγκεκριμένα, διαφορετικά μεγέθη παραθύρων χρησιμοποιούνται για τη δοκιμή της απόδοσης του διαγράμματος στα δεδομένα που έχουμε συλλέξει και, στη συνέχεια, το μέγεθος του παραθύρου που θα επιφέρει μια καλή ισορροπία μεταξύ του σφάλματος Τύπου I και Τύπου II, θα επιλεγεί. Όσο αφορά το κρύφο σώμα, το μέγεθός του είναι συνήθως μικρότερο από $2i+1$ και μεγαλύτερο από $i$, όπου $i$ το μέγεθος του στρώματος εισόδου, και τελικά προσδιορίζεται διαγράφοντας τους κόμβους στο κρυφό στρώμα από $2i+1$ σε $i$ για να επιτευχθεί η καλή απόδοση πρόβλεψης στο σύνολο των δεδομένων εκπαίδευσης. Το μέγεθος του στρώματος εξόδου έχει οριστεί να είναι 1.

Το σύνολο των δεδομένων αναπτύσσεται με βάση τη μέθοδο του κινούμενου παραθύρου. Το διάνυσμα παραθύρου που περιέχει $n$ σημεία δεδομένων $(x_{t,1}, x_{t,2}, ..., x_{t,n})$ τη στιγμή $t$ τροφοδοτείται στο RNN για να μάθει την αντιστοίχιση μεταξύ παραθύρου εισόδου και εξόδου $x_{t,n+1}$, στη φάση εκπαίδευσης. Για να προβλέψουμε το επόμενο σημείο $x_{t,n+1}$ στη φάση δοκιμής χρησιμοποιούμε την Εξίσωση (3.2). Η διαδικασία για τη ρύθμιση του διαγράμματος είναι η εξής:

1. Συλλέγουμε κανονικά δεδομένα από μια αυτοσυσχετιζόμενη διαδικασία.

2. Αναπτύσουμε το σύνολο των δεδομένων εκπαίδευσης με τη μέθοδο του κινούμενου

παραθύρου.

3. Ρυθμίζουμε τη δομή του δικτύου και τις παραμέτρους εκμάθησης.

4. Το σύνολο των δεδομένων εκπαίδευσης χρησιμοποιείται για την εκπαίδευση του μοντέλου RNN.

5. Όταν τα σφάλματα πρόβλεψης πληρούν την απαίτηση που δημιουργήθηκε στην φάση εκπαίδευσης, έχουμε κατασκευάσει ένα καλά εκπαιδευμένο μοντέλο RNN.

6. Οι τιμές πρόβλεψης συγκρίνονται με πραγματικές τιμές για τον υπολογισμό του σφάλματος πρόβλεψης.

7. Το ανώ όριο ελέγχου προσδιορίζεται με βάση το ARL = 370 (δηλαδή, σφάλμα τύπου I = 0.27%).

8. Τα δείγματα δοκιμής στη φάση παρακολούθησης τροφοδοτούνται στο καλά εκπαιδευμένο RNN για να ληφθεί το σφάλμα πρόβλεψης που θα σχεδιαστεί στο διάγραμμα ελέγχου.

Στην συνέχεια παραθέτουμε τα αποτελέσμτα των Chen και Yu (2019) για τις τιμές του ARL σε σχέση με το μέγεθος των παραθύρων για μια διαδικασία παραγωγής χαρτιού.

**Πίνακας 3.1:** Σύγκριση ARL για διαφορετικά μεγέθοι κινούμενων παραθύρων

| Window Size | 0.5 | 1.0 | 1.5 | 2.0 | 2.5 | 3.0 | 0.0 |
|---|---|---|---|---|---|---|---|
| 8 | 135.089 | 111.847 | 58.596 | 7.317 | 1.813 | 1.000 | 371.993 |
| 12 | 134.212 | 107.653 | 58.140 | 13.299 | 1.870 | 1.000 | 371.914 |
| 16 | 128.7750 | 105.79 | 56.455 | 13.556 | 2.093 | 1.056 | 371.336 |
| 20 | 127.680 | 101.770 | 54.544 | 15.407 | 2.178 | 1.125 | 369.982 |
| 24 | 123.351 | 98.258 | 39.567 | 14.269 | 2.877 | 1.150 | 372.247 |

## 3.4 Αποσυσχέτιση Δεδομένων

Σε αυτή την παράγραφο περιγράφουμε την διαδικασία που ακολουθούμε για την αποσυσχέτιση (data de-correlation) πολυδιάστατων συσχετισμένων δεδομένων. Υποθέτουμε πως

η εντός ελέγχου διαδικασία έχει μέσο $\mu$ και η χρονοσειρά των δεδομένων είναι στάσιμη με συνδιακύμανση $\gamma(s) = Cov(\mathbf{X}_i, \mathbf{X}_{i+s})$ για κάθε $i$ και $s$, και εξαρτάται μόνο από το $s$.

Η πρώτη παρατήρηση $\mathbf{X}_1$ έχει πίνακα συνδιακύμανσης $\gamma(0)$. Το τυποποιημένο διάνυσματά της ορίζεται ως:

$$\mathbf{X}_1^* = \gamma(0)^{-1/2}(\mathbf{X}_1 - \mu). \tag{3.8}$$

Στην συνέχεια, θεωρούμε το διάνυσμα $(\mathbf{X}_1^{'}, \mathbf{X}_2^{'})^{'}$, όπου $\mathbf{X}_2$ η δεύτερη παρατήρηση. Ο πίνακας συνδιακύμασης του γράφεται ως:

$$\Sigma_{2,2} = \begin{pmatrix} \gamma(0) & \sigma_1 \\ \sigma_1^{'} & \gamma(0) \end{pmatrix}, \tag{3.9}$$

όπου, $\sigma_1 = \gamma(1)$.
Τώρα, κάνοντας χρήση της παραγοντοποίησης Cholesky στον πίνακα $\Sigma_{2,2}$ έχουμε:

$$\Phi_2 \Sigma_{2,2} \Phi_2^{'} = \mathbf{D}_2,$$

όπου,

$$\Phi_{2,2} = \begin{pmatrix} \mathbf{I}_p & 0 \\ -\sigma_1^{'}\gamma(0)^{-1} & \mathbf{I}_p \end{pmatrix}, D_2 = \begin{pmatrix} \mathbf{d}_1 & 0 \\ 0 & \mathbf{d}_2 \end{pmatrix} = diag(\mathbf{d}_1, \mathbf{d}_2)$$

με $\mathbf{d}_1 = \gamma(0)$ και $\mathbf{d}_2 = \gamma(0) - \sigma_1^{'}\gamma(0)^{-1}\sigma_1$.
Επομένως, καταλήγουμε πως $Cov(\Phi_2 \mathbf{e}_2 = D_2)$, όπου $\mathbf{e}_2 = [(\mathbf{X}_1 - \mu)^{'}, (\mathbf{X}_2 - \mu)^{'}]^{'}$.
Αφού,

$$\Phi_2 \mathbf{e}_2 = \begin{pmatrix} \mathbf{I}_p & 0 \\ -\sigma_1^{'}\gamma(0)^{-1} & \mathbf{I}_p \end{pmatrix} \begin{pmatrix} (\mathbf{X}_1 - \mu)^{'} \\ (\mathbf{X}_2 - \mu)^{'} \end{pmatrix} = (\epsilon_1^{'}, \epsilon_2^{'})^{'}$$

με

$$\epsilon_1^{'} = \mathbf{X}_1 - \mu \tag{3.10}$$

$$\epsilon_2^{'} = -\sigma_1^{'}\Sigma_{1,1}^{-1}(\mathbf{X}_1 - \mu) + (\mathbf{X}_2 - \mu) \tag{3.11}$$

παρατηρούμε πως τα $\epsilon_1, \epsilon_2$ είναι ασυσχέτιστα. Είμαστε τώρα σε θέση να ορίσουμε το τυπο-

ποιημένο διάνυσμα του $\mathbf{X}_2$ το οποίο είναι:

$$\begin{aligned}\mathbf{X}_2^* &= \mathbf{d}_2^{-1/2}\epsilon_2 \\ &= \mathbf{d}_2^{-1/2}\big[-\sigma_1^{'}\Sigma_{1,1}^{-1}(\mathbf{X}_1-\mu)+(\mathbf{X}_2-\mu)\big]\end{aligned} \tag{3.12}$$

Τα $\mathbf{X}_1^*, \mathbf{X}_2^*$ είναι ασυσχέτιστα και έχουν πίνακα συνδιακύμανσης τον μοναδιαίο πίνακα $\mathbf{I}_p$. Ακολουθώντας την παραπάνω διαδικασία, μπορούμε να ορίσουμε τα ασυσχέτιστα και τυποποιημένα διανύσματα για κάθε παρατήρηση που συλλέγουμε. Πιο συγκεκριμένα, για την $j$-οστή παρατήρηση, ο πίνακας συνδιακύμανσης του διανύσματος $(\mathbf{X}_1^{'}, \mathbf{X}_2^{'}, ..., (\mathbf{X}_j^{'}))^{'}$ μπορεί να γραφεί ως:

$$\Sigma_{j,j} = \begin{pmatrix}\Sigma_{j-1,j-1} & \sigma_{j-1} \\ \sigma_{j-1}^{'} & \gamma(0)\end{pmatrix} \tag{3.13}$$

με $\sigma_{j-1} = ([\gamma(j-1)]^{'}, ..., [\gamma(2)]^{'}, [\gamma(1)]^{'})^{'}$.
Από παραγοντοποίηση Cholesky έχουμε:

$$\Phi_j\Sigma_{j,j}\Phi_j^{'} = \mathbf{D}_j \tag{3.14}$$

$$\Phi_j = \begin{pmatrix}\Phi_{j-1} & 0 \\ -\sigma_{j-1}^{'}\Sigma_{j-1,j-1}^{-1} & \mathbf{I}_p\end{pmatrix}, D_j = diag(\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_j) \tag{3.15}$$

με $\mathbf{d}_j = \Sigma_{j,j} - \sigma_{j-1}^{'}\Sigma_{j-1,j-1}^{-1}\sigma_{j-1}$.
Στην συνέχεια, ορίζουμε:

$$\epsilon_j^{'} = -\sigma_{j-1}^{'}\Sigma_{j-1,j-1}^{-1}\mathbf{e}_{j-1}(\mathbf{X}_j-\mu) \tag{3.16}$$

και τότε $\Phi_j\epsilon_j = (\mathbf{e}_1^{'}, \mathbf{e}_2^{'}, ..., \mathbf{e}_j^{'})^{'}$ και $Cov(\Phi_j\epsilon_j = D_j)$ από το οποίο συμπεραίνουμε πως το $\mathbf{e}_j^{'}$ είναι ασυσχέτιστο με το $(\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_{j-1})$. Επομένως το τυποποιμένο και ασυσχέτιστο διάνυσμα $\mathbf{X}_j$ είναι το:

$$\begin{aligned}\mathbf{X}_j^* &= \mathbf{d}_j^{-1/2}\epsilon_j \\ &= \mathbf{d}_j^{-1/2}\big[-\sigma_{j-1}^{'}\Sigma_{j-1,j-1}^{-1}\mathbf{e}_{j-1}(\mathbf{X}_j-\mu)\big]\end{aligned} \tag{3.17}$$

το οποίο είναι ασυσχέτιστο με τα $\mathbf{X}_1^*, ..., \mathbf{X}_{j-1}^*$ και έχει πίνακα συνδιακύμανσης $\mathbf{I}_p$.

Με την παραπάνω διαδικασία αποσυσχέτισης διαδοχικών δεδομένων, μπορούμε να

μετατρέψουμε τις αρχικά συσχετισμένες παρατηρήσεις διεργασίας σε μια ακολουθία ασυσχέτιστων και τυποποιημένων παρατηρήσεων, καθεμία από τις οποίες έχει το μέσο $\mathbf{0}$ και πίνακα συνδιακύμανσης $\mathbf{I}_p$. Στην πραγματικότητα, οι παράμετροι $\mu$ και $\gamma(s)$ είναι συνήθως άγνωστες και πρέπει να εκτιμηθούν εκ των προτέρων. Για το σκοπό αυτό, $\mu$ και $\gamma(s)$ μπορούν να εκτιμηθούν από το IC σύνολο δεδομένων $X_{IC} = (\mathbf{X_{m_0+1}}, \mathbf{X_{m_0+2}}, ..., \mathbf{X_0})$ ως εξής:

$$\hat{\mu} = \frac{1}{\mu_0} \sum_{i=-\mu_0+1}^{0} \mathbf{X}_i \tag{3.18}$$

$$\hat{\gamma} = \frac{1}{\mu_0 - s} \sum_{i=-\mu_0+1}^{-s} (\mathbf{X}_{i+s} - \hat{\mu})(\mathbf{X}_i - \hat{\mu})' \tag{3.19}$$

Για την παρακολούθηση μιας διαδικασίας με συσχετισμένες παρατηρήσεις, μπορούμε να αποσυσχετίσουμε διαδοχικά τις παρατηρήσεις χρησιμοποιώντας τη διαδικασία που περιγράφηκε παραπάνω και, στη συνέχεια, να εφαρμόσουμε τα διαγράμματα ελέγχου Μηχανικής Εκμάθησης που έχουμε αναφέρει. Ωστόσο, είναι πολύ πιθανό με αυτή την διαδικασία να χρειάζεται πολύς υπολογιστικός χρόνος, ειδικά όταν το τα δεδομένα μας είναι πάρα πολλά. Για τη μείωση του υπολογιστικού φόρτου, προτείνεται ότι η τελευταία παρατήρηση, έστω $\mathbf{X}_n$ χρειάζεται μόνο να αποσυσχετιστεί με τις προηγούμενες $b_{max}$ παρατηρήσεις. Βασιζόμαστε στην υπόθεση ότι δύο παρατηρήσεις γίνονται ασυσχετισμένες εάν οι χρόνοι παρατήρησής τους είναι μεγαλύτεροι από $b_{max}$. Αυτή η υπόθεση υποστηρίζει ότι η συσχέτιση δεδομένων είναι μικρής εμβέλειας, κάτι το οποίο είναι λογικό σε πολλές εφαρμογές. Με βάση αυτή την υπόθεση, συνοψίζουμε παρακάτω ένα τροποποιημένο διάγραμμα ελέγχου ML για την παρακολούθηση συσχετισμένων δεδομένων.

- Όταν $n = 1$, η ασυσχέτιστη και τυποποιημένη παρατήρηση ορίζεται ως:

$$\hat{\mathbf{X}}_1^* = \hat{\gamma}^{-1/2}(0)(\mathbf{X}_1 - \hat{\mu}). \tag{3.20}$$

  Θέτουμε την βοηθητική παράμετρος $b$ ίση με 1 και στην συνέχεια εφαρμόζουμε ένα ML διάγραμμα ελέγχου στο $\mathbf{X}_1^*$.

- Όταν $n > 1$, ο εκτιμώμενος πίνακας συνδιακύμανσης του $(\mathbf{X}_{n-b}^{'}, ..., \mathbf{X}_{n}^{'})$ ορίζεται ως:

$$\hat{\Sigma}_{n,n} = \begin{pmatrix} \hat{\gamma}(0) & ... & \hat{\gamma}(b) \\ \vdots & \ddots & \vdots \\ \hat{\gamma}(b) & ... & \hat{\gamma}(0) \end{pmatrix} =: \begin{pmatrix} \hat{\Sigma}_{n-1,n-1} & \hat{\sigma}_{n-1} \\ \hat{\sigma}_{n-1}^{'} & \hat{\gamma}(0) \end{pmatrix}.$$

Στην συνέχεια, η τυποποιημένη και ασυσχέτιστη παρατήρηση σε χρόνο $n$ ορίζεται ως:

$$\hat{\mathbf{X}}_{n}^{*} = \hat{\mathbf{d}}_{n}^{-1/2}\big[ - \hat{\sigma}_{n-1}^{'}\hat{\Sigma}_{n-1,n-1}^{-1}\hat{\mathbf{e}}_{n-1}(\mathbf{X}_{n} - \hat{\mu})\big]$$

όπου $\hat{\mathbf{d}}_{j} = \hat{\Sigma}_{j,j} - \hat{\sigma}_{j-1,j-1}^{'}\hat{\Sigma}_{n-1,n-1}^{-1}\hat{\sigma}_{j-1}$ και $\hat{\mathbf{e}}_{n-1} = [(\mathbf{X}_{n-b} - \hat{\mu})^{'}, (\mathbf{X}_{n-b+1} - \hat{\mu})^{'}, ..., (\mathbf{X}_{n-1} - \hat{\mu})^{'}]^{'}$. Εφαρμόζουμε ένα ML διάγραμμα ελέγχου στο $\hat{\mathbf{X}}_{n}^{*}$ για να δούμε εάν μας δώσει κάποιο σήμα. Αν όχι, θέτουμε $b = min(b+1, b_{max})$ και $n = n+1$ και παρακολουθούμε ξανά την διαδικασία.

Στην συνέχεια παραθέτουμε αποτελέσματα από τις συγκρίσεις των Xie και Qiu (2022) στα διαγράμματα ελέγχου RF, RTC, DSVM, KNN πριν και μετά την τροποποίηση για την αποσυσχέτιση που περιγράψαμε παραπάνω (RF-D, RTC-D, DSVM-D, KNN-D)

**Σχήμα 3.6:** Οι πραγματικές τιμές $ARL_0$ και τα τυπικά τους σφάλματα (σε παρένθεση)

| Methods | Case I | Case II | Case III | Case IV | Case V |
|---|---|---|---|---|---|
| RF | 189(3.98) | 194(4.20) | 105(1.42) | 119(2.05) | 106(1.33) |
| RF-D | 193(3.22) | 182(3.49) | 188(3.61) | 193(3.70) | 194(3.37) |
| RTC | 203(4.66) | 207(5.23) | 252(5.97) | 133(3.02) | 269(6.01) |
| RTC-D | 194(3.68) | 196(3.64) | 201(4.00) | 188(3.49) | 190(3.96) |
| DSVM | 213(5.20) | 195(4.77) | 263(6.99) | 118(2.87) | 277(6.34) |
| DSVM-D | 193(4.33) | 198(3.50) | 193(4.16) | 190(3.72) | 188(3.73) |
| KNN | 196(4.77) | 188(3.88) | 156(3.70) | 266(6.02) | 134(4.03) |
| KNN-D | 191(4.20) | 194(3.69) | 194(4.01) | 187(3.20) | 190(3.18) |

# Κεφάλαιο 4

# Διαγράμματα Ελέγχου με βάση τον Πυρήνα

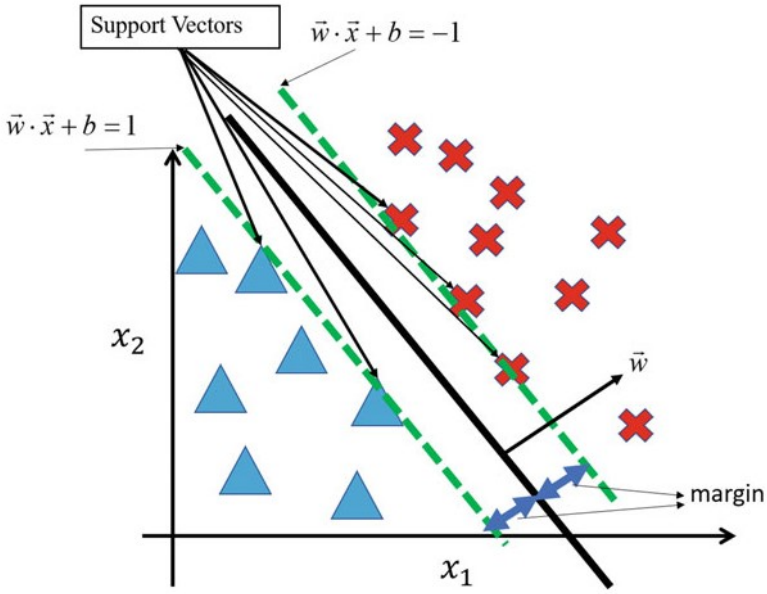
Σε αυτό το κεφάλαιο επικεντρωνόμαστε στα διαγράμματα ελέγχου που έχουν ως βάση τους τον πυρήνα (Kernel-distance-based control charts). Οι τεχνικές πυρήνα κερδίζουν συνεχώς έδαφος στον τομέα της παρακολούθησης διεργασιών λόγω των εντυπωσιακών αποτελεσμάτων τους. Υπάρχουν πολλές τεχνικές που χρησιμοποιούν (ή μπορούν να τροποποιηθούν άμεσα για χρήση) έναν πυρήνα. Αυτές οι τεχνικές, λαμβάνοντας τα πλεονεκτήματα των μη γραμμικών μετασχηματισμών που προκαλούνται από τους πυρήνες, μπορούν να αντιμετωπίσουν πιο δύσκολα προβλήματα. Σε πολλές περιπτώσεις, τα πρόβλημα παρακολούθησης διεργασιών μετατρέπονται σε προβλήματα παλινδρόμησης ή ταξινόμησης, για τα οποία ακριβώς σχεδιάστηκαν αρχικά οι πυρήνες. Η χρήση πυρήνων σε υπάρχουσες τεχνικές παρακολούθησης χρησιμοποιούνται εκτενώς και διαδραματίζουν σημαντικό ρόλο στην απόδοση των μεθόδων. Μάλιστα, σε πολλές περιπτώσεις, οι ιδιότητες των πυρήνων αποτελούν το βασικό στοιχείο μιας επιτυχής επίλυση των πραγματικών προβλημάτων. Αυτός είναι ο κύριος λόγος για τον οποίο οι μέθοδοι πυρήνα έχουν γίνει τόσο σημαντικές σε αυτό το πεδίο.

## 4.1 Μέθοδοι Μάθησης με βάση τον πυρήνα

Οι μέθοδοι μάθησης με κέντρο τον πυρήνα (kernel-based learning methods) χρησιμοποιούνται ευρέως στον στατιστικό έλεγχο διεργασιών. Διαδραματίζουν σημαντικό ρόλο στην σχεδίαση διαγραμμάτων ελέγχου και στην αναγνώριση ανωμαλιών λόγω των σπουδαίων λύσεων που παρέχουν. Η εφαρμογή των πυρήνων είναι πολύ σημαντική. Πολλές φορές είναι δύσκολο να εργαστούμε στον αρχικό χώρο του προβλήματος και η μέθοδος του πυρήνα μας δίνει τη δυνατότητα να μετατρέψουμε τον αρχικό χώρο σε έναν άλλον, στον οποίο μπορούμε να εργαστούμε ευκολότερα.

Οι αλγόριθμοι που βασίζονται στον πυρήνα χρησιμοποιούνται κυρίως σαν ταξινομητές (classifiers), δηλαδή χωρίζουν τα δεδομένα σε 2 ή περισσότερες κλάσεις (classes). Στον ΣΕΔ χρησιμοποιούνται οι δύο εξής κλάσεις: η κανονική κατάσταση (normal state) και η μη κανονική κατάσταση (abnormal state). Σε πολλές περιπτώσεις όμως, δεν έχουμε στην διάθεσή μας δεδομένα και για τις δύο κλάσεις, κυρίως για την μη κανονική κατάσταση (Weese M., Martinez W., Megahed F. M., Jones-Farmer L. A. (2016)). Για αυτόν τον λόγο εισάγουμε τους ταξινομητές μία κλάσης (one-class classifiers). Οι ταξινομητές αυτοί μαθαίνουν από τα κανονικά δεδομένα εκπαίδευσης και χαρακτηρίζουν τα δεδομένα που προκύπτουν ως παρατηρήσεις εντός κλάσης ή παρατηρήσεις εκτός κλάσης.

Ένας αλγόριθμος που ανήκει στην κατηγορία των μεθόδων με βάση των πυρήνα είναι αυτός των Μηχανών Διανυσμάτων Υποστήριξεις (Support Vector Machine - SVM). Οι SVM είναι νέες στατιστικές τεχνικές μάθησης. Βοηθούν στην αντιμετώπιση διαφόρων θεμάτων όπως την ταξινόμηση, την παλινδρόμηση, την σύντηξη κ.λ.π.. Η βασική ιδέα των διανυσμάτων υποστήριξης συνίσταται στην προβολή των δεδομένων του χώρου εισόδου, που είναι μη γραμμικά διαχωρίσιμα (ανήκουν σε δύο διαφορετικές κλάσεις), σε έναν χώρο μεγαλύτερης διάστασης, που ονομάζεται χώρος των χαρακτηριστικών (space of characteristics), με τέτοιο τρόπο ώστε τα δεδομένα να γίνονται γραμμικώς διαχωρίσιμα. Σε αυτόν τον χώρο, χρησιμοποιούμε την τεχνική κατασκευής του βέλτιστου υπερεπιπέδου για τον υπολογισμό της συνάρτησης ταξινόμησης που διαχωρίζει τα δεδομένα σε δύο κλάσεις. Δηλαδή, ο αλγόριθμος δημιουργεί μια γραμμή ή ένα υπερεπίπεδο, που διαχωρίζει τα δεδομένα σε κλάσεις.




**Σχήμα 4.1:** Τεχνική Διανύσματος Υποστήριξης

Κάποιοι από τους βασικότερους ταξινομητές μιας κλάσης που χρησιμοποιούνται στα διαγράμματα ελέγχου είναι οι εξής: Διάνυσμα Υποστήριξης για περιγραφή δεδομένων (support vector data description) και $k$ κοντινότερος γείτονας ($k$ nearest neighbor).

### 4.1.1 Διάνυσμα υποστήριξης για περιγραφή δεδομένων

Είναι μια μέθοδος που ενσωματώνει τις Μηχανές Διανυσμάτων Υποστήριξης και τον αλγόριθμο περιγραφής δεδομένων. Αφού έχουμε ταξινόμηση μίας τάξης, το υπερεπίπεδο της μεθόδου των Μηχανών Διανυσμάτων Υποστήριξης, γίνεται μια ενιαία υπερσφαίρα, η οποία περικλείει μια επιθυμητή αναλογία παρατηρήσεων. Το όριο απόφασης του Διανύσματος Υποστήριξης είναι κατασκευασμένο ακριβώς για να ελαχιστοποιεί τον όγκο της υπερσφαίρας μεγιστοποιώντας των αριθμό των παρατηρήσεων εκπαίδευσης που περιέχονται στην υπερσφαίρα (Lee S., Lee S., Kim C.K. (2022)). Στόχος είναι επίσης η εύρεση του κέντρου και της ακτίνας της. Το όριο απόφασης είναι κατασκευασμένο ακριβώς για να ελαχιστοποιεί τον όγκο της υπερσφαίρας μεγιστοποιώντας ταυτόχρονα τον αριθμός των εκπαιδευτικών

παρατηρήσεων που περιέχονται στην υπερσφαίρα. Συγκεκριμένα, έστω $\alpha$ το κέντρο της υπερσφαίαρας και $R$ η ακτίνα της. Ακόμη, έστω $x_i$, $i = 1, ..., n$ τα διανύσματα του συνόλου δεδομένων εκπαίδευσης. Για να λάβουμε τα βέλτιστα $\alpha$ και $R$ έχουμε το πρόβλημα βελτιστοποίησης:

$$R^2 + C\sum_{i=1}^{n}\xi_i \tag{4.1}$$

με τον εξής περιορισμό

$$\|x_i - a\|^2 \leq R^2 + \xi_i, \tag{4.2}$$

όπου $\xi_i$ η slack μεταβλητή και $C$ η παράμετρος συντονισμού που ελέγχει την ισορροπία μεταξύ του όγκου της υπερσφαίρας και των σφαλμάτων ταξινόμησης.
Ενσωματώνοντας τους περιορισμούς (4.2) στη (4.1) κατασκευάζουμε την Λαγκραντζιανή ως εξής:

$$\begin{aligned} L &= L(R, \alpha, \alpha_i, \gamma_i, \xi_i) \\ &= R^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[R^2 + \xi_i - (\|x_i - \alpha\|^2)] - \sum_{i=1}^{n}\gamma_i\xi_i, \end{aligned} \tag{4.3}$$

όπου $\alpha_i, \gamma_i \geq 0$ οι πολλαπλασιαστές Lagrange. Μηδενίζοντας τις μερικές παραγώγους της Λαγκραντζιανής προκύπτουν οι νέοι περιορισμοί:

$$\sum_{i=1}^{n}\alpha_i = 1, \alpha = \sum_{i=1}^{n}\alpha_i x_i, \alpha_i = C - \gamma_i \tag{4.4}$$

Επειδή $\alpha_i, \gamma_i \geq 0$, απαλλοίφονται οι μεταβλητές $\gamma_i$ από την τρίτη Εξίσωση των (4.4) και να χρησιμοποιηθούν οι περιορισμοί $0 \leq \alpha_i \leq C$. Ξαναγράφοντας την Εξίσωση (4.3) και αντικαθιστώντας σε αυτή τους νέους περιορισμούς το πρόβλημα ανάγεται στην:

$$L = \sum_{i=1}^{n} a_i\langle x_i, x_j\rangle - \sum_{i,j=1}^{n} a_i a_j\langle x_i, x_j\rangle \tag{4.5}$$

και η λύση των $\alpha_i$ λαμβάνεται από την μεγιστοποίηση της παραπάνω εξίσωσης. Στην συνέχεια με την δεύτερη Εξίσωση των (4.4) μπορούμε να εκτιμήσουμε το κέντρο $\alpha$. Το κέντρο της υπερσφαίρας είναι ένας γραμμικός συνδυασμός των αντικειμένων με παράγοντες βάρους $\alpha_i$. Για αυτά τα αντικείμενα, που οι αντίστοιχοι συντελεστές τους $\alpha_i$ είναι διάφο-

ροι του μηδενός, ονομάζονται διανύσματα υποστήριξης (Support Vectors). Μόνο αυτά τα αντικείμενα είναι απαραίτητα για τη περιγραφή της υπερσφαίρας. Η ακτίνα της υπερσφαίρας μπορεί να προκύψει από τον υπολογισμό της απόστασης του κέντρου της από ένα διάνυσμα υποστήριξης με συντελεστή $\alpha_i < C$. Αντικείμενα για τα οποία $\alpha_i = C$, βρίσκονται έξω από τη σφαίρα. Για να βρούμε ένα όριο απόστασης έχουμε την εξής διαδικασία: Αρχικά, αντικαθιστούμε το εσωτερικό γινόμενο στην εξίσωση (4.5) με:

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \tag{4.6}$$

όπου $\phi(x)$ είναι ένας έμμεσος τελεστής που κατασκευάζει τον πυρήνα. Συνήθως χρησιμοποιούμε την Γκαουσιανή συνάρτηση πυρήνα

$$K(x_i, x_j) = exp\left( - \frac{\|x_i - x_j\|}{\kappa} \right) \tag{4.7}$$

όπου $\kappa > 0$ είναι μια παράμετρος συντονισμού που ελέγχει το πλάτος του πυρήνα. Χρησιμοποιούμε λοιπόν την Γκαουσιανή συνάρτηση και υπολογίζουμε την απόσταση ανάμεσα σε μια παρατήρηση $z$ και του κέντρου $\alpha$ ως εξής:

$$D^2 := D^2(z) = K(z, z) - 2 \sum_{i=1}^{n} a_i K(z, x_i) + \sum_{i,j=1}^{n} a_i a_j K(x_i, x_j) \tag{4.8}$$

και το $z$ θεωρείται εκτός της υπερσφαίρας όταν το $D^2$ είναι μεγαλύτερο του $R^2$. Το τετράγωνο της ακτίνας της υπερσφαίρας είναι πολύ σημαντικό, καθώς παίζει κρίσιμο ρόλο στον καθορισμό των ορίων ελέγχου των διαγραμμάτων ελέγχου.

### 4.1.2 k κοντινότερος γείτονας (k-nearest neighbor - KNN)

Είναι μια από τις απλούστερες μεθόδους αναγνώρισης μοτίβων, η οποία ταξινομεί σύμφωνα με μια δοσμένη τιμή $k$ από την κλάση του πλησιέστερου γείτονα. Είναι μια μη παραμετρική μέθοδος ταξινόμησης επιβλεπόμενης μάθησης. Σε αντίθεση με άλλες επιβλεπόμενες τεχνικές εκμάθησης, δεν έχει φάση εκπαίδευσης. Οι κλάσεις σε ένα σύνολο δεδομένων καθορίζονται από τα ιστορικά δεδομένα. Κάθε δείγμα στο σύνολο δεδομένων που πρόκειται να ταξινομηθεί στη δοκιμαστική φάση υποβάλλεται σε επεξεργασία ξεχωριστά. Για να προσδιοριστεί η κλάση

αυτού του δείγματος, πρέπει να επιλεγεί ο αριθμός $k$, ο οποίος είναι ο αριθμός των γειτόνων του άγνωστου δείγματος. Οι $k$-πλησιέστεροι γείτονες καθορίζονται με βάση ορισμένες συναρτήσεις απόστασης όπως η Ευκλείδεια απόσταση. Η συγκεκριμένη γενικά προτιμάται λόγω της καταλληλότητας της για την κατανομή Gauss και της ευκολίας χρήσης της. Η συνάρτηση της απόστασης απο το $p = (p_1, p_2, ..., p_n)$ στο $q = (q_1, q_2, ..., q_n)$ είναι:

$$EUD = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

### 4.1.3 Τεχνική Πυρήνα Naive Bayes

Ο πυρήνας του Naive Bayes kernel (NB-k) είναι ένας απλός ταξινομητής και βασίζεται στο θεώρημα Bayes. Προϋποθέτει ανεξαρτησία μεταξύ κάθε τάξης. Οι τεχνικές NB-k καθορίζουν την δεσμευμένη πιθανότητα για την σχέση μεταξύ κάθε μεταβλητής και κλάσης για κάθε στοιχείο (Demircioglu Diren D., Boran S., Cil I. (2020)). Χρησιμοποιείται όταν ο αριθμός των μεταβλητών είναι μεγάλος. Το θεώρημα του Bayes παρουσιάζεται ως εξής:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \tag{4.9}$$

όπου $P(C|X)$ η posterior πιθανότητα, $P(X|C)$ η πιθανότητα του $X$ δεδομένου του $C$ και $P(C)$ η πιθανότητα να λάβουμε την τάξη.

Έχει παρατηρηθεί ότι τα ποσοστά ακρίβειας της τεχνικής Naive Bayes είναι μεγαλύτερα όταν χρησιμοποιείται μαζί με μία συνάρτηση πυκνότητας πυρήνα (Murakami και Mizuguchi (2010)). Στη μέθοδο της συνάρτησης πυκνότητας πυρήνα, το εύρος ζώνης προσδιορίζεται αφού επιλεγεί ο αριθμός του πυρήνα. Υπάρχουν πολλές διαφορετικές μέθοδοι για τον προσδιορισμό του εύρους ζώνης. Διαφορετικά, καθορίζεται από ειδικούς (Kuter, Usul και Kuter (2011)).

## 4.2 Διαγράμματα Ελέγχου με βάση τον πυρήνα

Οι μέθοδοι πυρήνα μεταφέρουν τα δεδομένα από τον αρχικό χώρο όπου βρίσκονται, γνωστός ως χώρος εισόδου, σε έναν χώρο υψηλότερων διαστάσεων, γνωστός ως χώρος χαρακτηριστικών. Στη συνέχεια, αυτές οι μέθοδοι αναζητούν συναρτήσεις γραμμικής απόφασης στον χώρο των χαρακτηριστικών οι οποίες γίνονται μη γραμμικές συναρτήσεις απόφασης στον χώρο εισόδου. Για να γίνει αυτό, οι μέθοδοι πυρήνα αντικαθιστούν το εσωτερικό προϊόν των παρατηρήσεων με μια επιλεγμένη συνάρτηση πυρήνα. Εάν το $h(\cdot)$ είναι ένας μετασχηματισμός από το χώρο εισόδου $\Re^p$ στον χώρο χαρακτηριστικών $F$, μια συνάρτηση πυρήνα ορίζεται ως:

$$K(x, y) = \langle h(x), h(y) \rangle_F$$

για $x, y \in R^p$. Το θετικό είναι ότι δεν χρειάζεται να γνωρίζουμε την ακριβή μορφή του $h(\cdot)$. Ένας από τους πιο συχνά χρησιμοποιούμενους πυρήνες είναι η Γκαουσιανή ακτινική συνάρτηση βάσης, που ορίζεται ως εξής:

$$K(x, y) = exp(-\gamma \|x - y\|^2)$$

με $\gamma > 0$. Άλλες διάσημες συναρτήσεις είναι οι εξής:

- Η γραμμική $K(x, y) = x^{'} y$.
- Η πολυωνυμική $K(x, y) = (\gamma x^{'} y + c)^d$.
- Η Sigmoid $K(x, y) = tanh(\gamma x^{'} y + c)^d)$.

όπου $\gamma > 0$, $c, d$ παράμετροι των συναρτήσεων πυρήνα. Αυτές είναι οι κύριες συναρτήσεις που χρησιμοποιούνται και στα διαγράμματα που θα αναλύσουμε στην συνέχεια αυτού του Κεφαλαίου.

### 4.2.1 Διάγραμμα Ελέγχου με χρήση του Αλγορίθμου SVM

Το RTC διάγραμμα χρησιμοποιεί διακριτές τιμές καθώς τα $P_{OOB}(\mathbf{X}_i)$ υπολογίζονται από τις αποφάσεις των δέντρων. Μια εντελώς διαφορετική προσέγγιση είναι αυτή της δημιουρ-

γίας ενός διαγράμματος που χρησιμοποιεί την απόσταση με την επιλογή του αλγορίθμου SVM. Το συγκεκριμένο διάγραμμα ονομάζεται DSVM (Distance SVM) και χρησιμοποιεί την απόσταση μεταξύ των διανυσμάτων υποστήριξης και των παρατηρήσεων της διαδικασίας $X_{AC_n}$. Η απόσταση από ένα δείγμα των παρατηρήσεων μέχρι την οριακή επιφάνεια, που ορίζεται από το διάνυσμα υποστήριξης, μπορεί να είναι είτε θετική είτε αρνητική. Για αυτόν τον λόγο, μετατρέπουμε την απόσταση με την χρήση της τυπικής λογιστικής συνάντησης

$$g(a) = \frac{1}{1 + exp(-\alpha)}. \tag{4.10}$$

Η μέση τιμή των μετασχηματισμένων αποστάσεων από μεμονωμένες παρατηρήσεις στο $X_{AC_n}$ στην οριακή επιφάνεια μπορεί να οριστεί ως το στατιστικό:

$$M_n = \frac{\sum g(d(\mathbf{X}_i))I(\mathbf{X_i} \in X_{AC_n})}{|X_{AC_n}|}, \tag{4.11}$$

όπου $d(\mathbf{X}_i)$ είναι η απόσταση από την παρατήρηση $\mathbf{X}_i$ στο όριο που καθόρισε ο αλγόριθμος SVM σε χρόνο $n$. Μένει να επιλέξουμε την συνάρτηση πυρήνα και την παράμετρο ποινής. Στο μοντέλο DSVM χρησιμοποιείται η Γκαουσιανή συνάρτηση πυρήνα

$$K(\mathbf{X}, \mathbf{X}') = exp\left(\frac{\langle \mathbf{X} - \mathbf{X}' \rangle^2}{\sigma^2}\right) \tag{4.12}$$

για κάθε $\mathbf{X}, \mathbf{X}' \in R^p$, όπου $p$ η διάσταση των παρατηρήσεων της διαδικασίας, και η παράμετρος $\sigma^2$ επιλέχτηκε να είναι μεγαλύτερη του 2.8. Επίσης, η παράμετρος ποινής αποφασίστηκε να είναι ίση με 1. Τα όρια ελέγχου καθορίζονται παρόμοια με αυτά του διαγράμματος RTC.

### 4.2.2 Διάγραμμα Ελέγχου με ταξινομητή KNN

Μία άλλη προσέγγιση στην σχεδίαση διαγραμμάτων ελέγχου είναι η χρήση των ταξινομητών μίας τάξης. Αρχικά, αναπτύχθηκε ένα μη παραμετρικό διάγραμμα ελέγχου που βασίζεται στα διανύσματα υποστήριξης (Support vector data description - SVDD). Με το SVDD η οριακή επιφάνεια των εντός ελέγχου δεδομένων ορίζεται έτσι ώστε ο όγκος εντός της οριακής επιφάνειας να είναι όσο το δυνατόν μικρότερος, ενώ το σφάλμα τύπου I διατηρείται εντός ενός δοσμένους επιπέδου $\alpha$. Μετά, η οριακή επιφάνεια θα χρησιμοποιείται ως κανόνας απόφασης

για την παρακολούθηση της διαδικασίας ως εξής: Μία παρατήρηση θεωρείται εκτός ελέγχου αν είναι έξω από το όριο και εντός ελέγχου στην άλλη περίπτωση. Όμως, ο υπολογισμός του ορίου με αυτόν τον τρόπο έχει μεγάλο υπολογιστικό κόστος. Για την μείωση του υπολογιστικού φόρτου προτείνεται ένα διάγραμμα ελέγχου βασισμένο στον ταξινομητή KNN. Σε αυτή την περίπτωση, η μέση απόσταση μεταξύ μιας δεδομένης δοσμένη παρατήρησης $X_i$ και των $k$-κοντινότερων γειτονικών παρατηρήσεων στο IC σύνολο δεδομένων υπολογίζεται αρχικά ως

$$K_i^2 = \frac{\sum_{j=1}^{k} \langle \mathbf{X}_i - NN_j(\mathbf{X}_i) \rangle}{k}, \tag{4.13}$$

όπου $NN_j(\mathbf{X}_i)$ είναι η $j$-οστή γειτονική παρατήρηση του $X_i$ στο IC σύνολο, και $\langle \bullet \rangle$ η ευκλείδια απόσταση. Η παρακολούθηση της διαδικασίας γίνεται ως εξής: Την στιγμή $n$, αν η μέση απόσταση από το $X_n$ στην $k$-κοντινότερη γειτονική παρατήρηση είναι μικρότερη από το $(1-\alpha)$-ιοστό τεταρτημόριο, το $X_n$ είναι εντός των ορίων. Σε αντίθετη περίπτωση είναι εκτός. Η εύρεση των ορίων ελέγχου ακολουθεί μία bootstrap διαδικασία. Αρχικά, ένα σύνολο από $B = 1000$ bootstrap δείγματα επιλέγονται από το σύνολο δεδομένων IC με τυχαία επιλογή. Μετέπειτα, υπολογίζεται το $(1-\alpha)$-ιοστό τεταρτημόριο του $\{K_i^2\}$ των μεμονομένων παρατηρήσεων σε κάθε bootstrap δείγμα. Τέλος, το τελικό όριο ελέγχου επιλέγεται ως το μέσο των $B$ τέτοιων τεταρτημορίων. Υποθέτουμε καθόλη την διαδικασία ότι οι παρατηρήσεις είναι ανεξάρτητες. Επομένως, το $ARL_0$ ισούται με $1/\alpha$.

### 4.2.3 $K$ Πολυμεταβλητό Διάγραμμα Ελέγχου με βάση τον πυρήνα

Ένα σημαντικό διάγραμμα που βασίζεται στο SVM είναι το Πολυμεταβλητό Διάγραμμα Ελέγχου με βάση τον πυρήνα (kernel-distance-based multivariate control chart), γνωστό και ως $K$-διάγραμμα (K-chart). Η ιδέα του συγκεκριμένου διαγράμματος είναι πως κατά την παρακολούθηση δύο ανεξάρτητων μεταβλητών μετά από διμεταβλητή κανονική κατανομή, το όριο ελέγχου διαμορφώνεται ως έλλειψη με κάθετους ή οριζόντιους άξονες (Sun και Tsung (2003)). Γενικότερα, σε προβλήματα της πραγματικής ζωής, το όριο ελέγχου μπορεί να μην είναι υποχρεωτικά ελλειπτικό, αλλά μπορεί να προσαρμοστεί σε πραγματικά δεδομένα, δίνοντας μια πιο ευέλικτη μορφή, προσαρμοσμένη σε οποιαδήποτε κατάσταση. Με

βάση αυτή την ιδέα, το $K$-διάγραμμα μπορεί να εκμεταλλευτεί τις αρχές του SVM, χρησιμοποιώντας διανύσματα υποστήριξης για την κατασκευή του ορίου ελέγχου. Εξ ορισμού, το $K$-διάγραμμα δίνει τον ελάχιστο όγκο ενός κλειστού σφαιρικού ορίου γύρω από τα εντός ελέγχου δεδομένα διεργασίας. Μετρά την απόσταση μεταξύ του κέντρου του πυρήνα και του νέου δείγματος για παρακολούθηση, το οποίο μπορεί να υπολογιστεί χρησιμοποιώντας διανύσματα υποστήριξης. Το $K$-διάγραμμα χρησιμοποιεί τη μέθοδο SVDD για την κατασκευή μιας σφαίρας που περιέχει το μέγιστο των δεδομένων. Οποιοδήποτε σημείο εκτός της σφαίρας θεωρείται εκτός ελέγχου.

Ο σχεδιασμός του $K$-διαγράμματος απαιτεί τον προσδιορισμό του κέντρου του πυρήνα και την κατασκευή του ορίου ελέγχου. Στην πραγματικότητα, ο στόχος του διαγράμματος είναι να βρεθεί το κέντρο της υπερσφαίρας που απαιτείται, για τον έλεγχο της κατάστασης των νέων δειγμάτων, λύνοντας τον παρακάτω τετραγωνικό προγραμματισμό:

$$Maximize : \sum_{i=1}^{1} \alpha_i K(x_i.x_i) - \sum_{i,j=1}^{1} \alpha_i \alpha_j K(x_i.x_j). \tag{4.14}$$

Με

$$\alpha_i \geq 0 \tag{4.15}$$

$$\sum_{i=1}^{1} \alpha_i = 1. \tag{4.16}$$

Πράγματι, για να προσδιοριστεί εάν ένα νέο δείγμα, έστω $z$, είναι εντός ελέγχου ή εκτός ελέγχου, θα πρέπει να υπολογιστεί μια απόσταση $d$ μεταξύ του $z$ και του $\alpha$, που είναι το κέντρο της υπερσφαίρας. Αυτό δίνεται από την ακόλουθη εξίσωση:

$$d = \sqrt{(z-a)'(z-a)}. \tag{4.17}$$

Εάν το $d$ είναι μεγαλύτερο από την ακτίνα $R$ της υπερσφαίρας, το δείγμα $z$ θεωρείται εκτός ελέγχου και εάν το $d$ είναι μικρότερο από το $R$, το δείγμα $z$ θεωρείται εντός ελέγχου.

Η απόσταση από τον πυρήνα (kernel distance - KD) μπορεί να υπολογιστεί με χρήση

της εξίσωσης:

$$KD = \sqrt{z'z - 2\sum_{i \in S} \alpha_i K(z.x_i) + \sum_{i,j \in S} \alpha_i, \alpha_j K(x_i.x_j)}, \tag{4.18}$$

όπου $S$ το σύνολο των αριθμών που αντιστοιχούν στα διανύσματα υποστήριξης. Το $KD$ αντιπροσωπεύει το ανώτερο όριο ελέγχου (upper control limit - UCL) για το $K$-διάγραμμα. Αυτό μπορεί να απεικονιστεί με το ακόλουθο τεστ υποθέσεων:

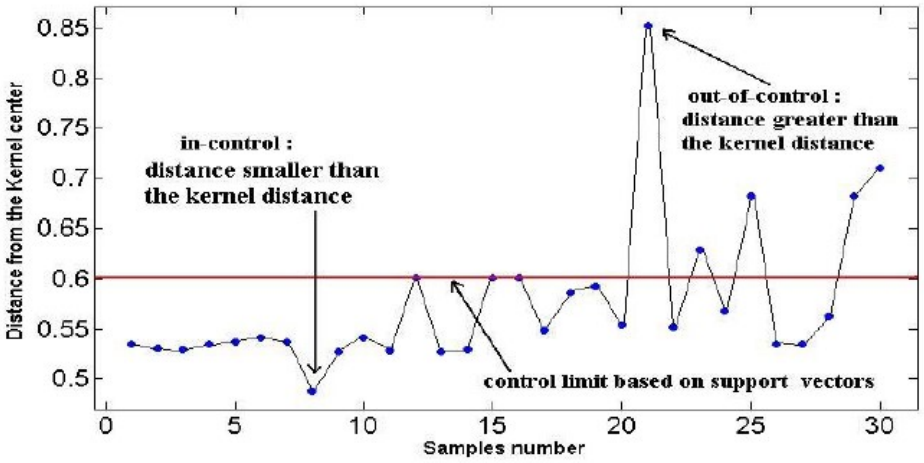
$$H_0 : distance(z) = KD \qquad vs \qquad H_1 : distance(z) > KD \tag{4.19}$$

από όπου το δείγμα $z$ θεωρείται εντός ελέγχου κάτω από την υπόθεση $H_0$ και εκτός ελέγχου κάτω από την υπόθεση $H_0$.

Στη φάση I, το σύνολο δεδομένων χρησιμοποιείται ως σύνολο εκπαίδευσης για τον προσδιορισμό της απόστασης του πυρήνα και την ανίχνευση ακραίων τιμών, για την κατασκευή του ορίου που περιβάλλει τα δεδομένα. Σε αυτή την φάση, θα πρέπει να επαληθεύσουμε ότι τα δεδομένα είναι ανεξάρτητα και ακολουθούν την ίδια κατανομή, καθώς αυτό είναι η θεμελιώδης υπόθεση του SVM. Στη φάση II, το $K$-διάγραμμα καθορίζει εάν τα νέα δείγματα είναι εντός ή εκτός ελέγχου συγκρίνοντας την απόσταση κάθε δείγματος με την απόσταση του πυρήνα. Οποιοδήποτε δείγμα έχει απόσταση μεγαλύτερη από την απόσταση του πυρήνα θα θεωρείται εκτός ελέγχου και αντίστροφα. Μια σημαντική υπόθεση σχετικά με τα δεδομένα είναι ότι πρέπει να είναι ανεξάρτητα και ομοίως κατανεμημένα. Επιπλέον, το όριο ελέγχου βασίζεται σε διανύσματα υποστήριξης. Επομένως, είναι κατάλληλο για πραγματικά δεδομένα και μπορεί να λάβει οποιαδήποτε μορφή ανάλογα με την περίπτωση μας. Η κεντρική γραμμή του διαγράμματος αντιπροσωπεύεται από το κέντρο του πυρήνα. Επιπλέον, δεν υπάρχουν περιορισμένες υποθέσεις σχετικά με τη κατανομή των διεργασιών, καθώς το $K$-διάγραμμα δέχεται οποιαδήποτε γνωστή ή άγνωστη υπόθεση.

### 4.2.4 $KM$ Διάγραμμα Ελέγχου

Ένα ακόμα διάγραμμα ελέγχου με βάση τον πυρήνα, που προτάθηκε από τους Gani και Limam M.(2014), είναι το διάγραμμα $KM$. Αποτελεί ένα διάγραμμα ελέγχου μιας κλάσης




**Σχήμα 4.2:** Μορφή $K$-διαγράμματος

της κατηγορίας της μη επιβλεπόμενης μάθησης που βασίζεται στον αλγόριθμο $K-means$. Δίνει το ελάχιστο κλειστό σφαιρικό όριο γύρω από τα εντός ελέγχου δεδομένα με χρήση του αλγορίθμου KMDD. Μετράει, δηλαδή, την απόσταση μεταξύ της σφαίρα του αλγορίθμου KMDD και του νέου δείγματος που εισέρχεται για παρακολούθηση. Αυτή η σφαίρα αποτελείται από $K$ συστάδες, οι οποίες είναι τοποθετημένες με τέτοιο τρόπο ώστε η μέση απόσταση του κέντρου μιας συστάδας να ελαχιστοποιείται.

Αρχικά, στην φάση I προσδιορίζεται η βέλτιστη μια κλάση του αλγορίθμου KMDD. Πρέπει να εκτιμηθεί ο βέλτιστος αριθμός συστάδων. Ο αλγόριθμος $K-means$ προσπαθεί να υπολογίσει $K$ συστάδες, έστω $C_1, ..., C_K$ που ελαχιστοποιούν το άθροισμα τετραγώνων εντός των συστάδων ως εξής:

$$D^2 = \min_{C_1,...C_K} \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - \mu_k\|^2 \tag{4.20}$$

με $i = 1, .., n, k = 1, ..., K$, όπου $C_1, ..., C_K$ είναι οι συστάδες, $n$ ο αριθμός των παρατηρήσεων στην φάση εκπαίδευσης, $\mu_k$ ο δειγματικός μέσος των παρατηρήσεων στην $k$-συστάδα και $D^2$ η Ευκλείδια απόσταση του ποιοτικού χαρακτηριστικού $x_i$. Οι τιμές του $D^2$ είναι αυτές που χρησιμοποιούνται στην κατασκευή του διαγράμματος ελέγχου. Το πρόβλημα βελτιστοποίησης (4.20) λύνεται με τον εξής τρόπο: Έστω οι κεντρικές συστάδες $\mu_1, ..., \mu_K$, αντιστοιχούν κάθε σημείο στην συστάδα με το κοντινότερο κέντρο. Μετά από κάθε αντιστοίχηση συστάδας, υπολογίζουμε ξανά τα κέντρα των συστάδων ώστε να είναι ο δειγματικός μέσος

των παρατηρήσεων σε κάθε συστάδα.

Στην Φάση ΙΙ, η απόσταση ενός νέου δείγματος, έστω $z_j$, ορίζεται ως:

$$D^2(z_j) = \min_{1,...,k} \|z_j - \mu_k\|^2, \quad j = 1, ..., m, k = 1, ..., K, \tag{4.21}$$

όπου $m$ είναι ο αριθμός των παρατηρήσεων στην φάση εκπαίδευσης.

Το $z_j$ είναι εντός ελέγχου όταν η απόστασή του είναι μικρότερη ή ίση με την ακτίνα της σφαίρας του KMDD, που ορίζεται ως $R^2_{KMDD}$. Δηλαδή:

$$z_j = \begin{cases} \text{εντός ελέγχου}, & \text{αν } D^2(z_j) \leq R^2_{KMDD} \\ \text{εκτός ελέγχου}, & \text{αν } D^2(z_j) \geq R^2_{KMDD} \end{cases}$$

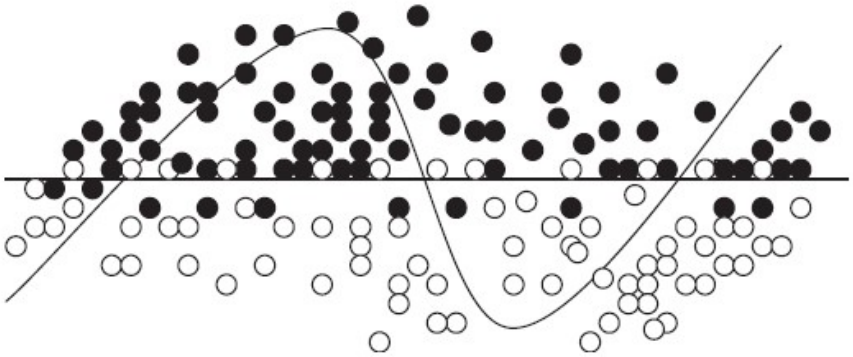
Η ακτίνα $R^2_{KMDD}$ αποτελεί την ακτίνα του ταξινομητή μίας κλάσης που βασίζεται στο KMDD και αντιπροσωπεύει και το άνω όριο ελέγχου (UCL) στο διάγραμμα KM. Η τιμή του επιρρεάζεται από τον αριθμό των συστάδων που χρησιμοποιούνται για την κατασκευή της κλάσης.

### 4.2.5 Ισχυρό $k$-Διάγραμμα με βάση το RSVM

Οι Μηχανές Διανυσμάτων Υποστήριξης (SVM) αποτελούν ένα πολύ σημαντικό εργαλείο στην ταξινόμηση και χρησιμοποιούνται ευρέως σε πολλά προβλήματα. Μία από τις κύριες υποθέσεις που κάνουμε όταν χρησιμοποιούμε SVM είναι πως όλα τα δείγματα στο σετ εκπαίδευσης είναι ανεξάρτητα και όμοια κατανεμημένα. Ωστόσο, τα πραγματικά δεδομένα που λαμβάνονται συχνά επιρρεάζονται από τον θόρυβο (noise). Επιπλέον, μερικά δείγματα στο σύνολο εκπαίδευσης τοποθετούνται στη λάθος πλευρά. Τέτοια σύνολα δεδομένων ονομάζονται ακραίες τιμές (outliers). Στις αυτές τις περιπτώσεις, όταν χρησιμοποιείται ένας τυπικός αλγόριθμος SVM, το όριο απόφασης αποκλίνει σοβαρά από το βέλτιστο υπερεπίπεδο. Αυτό σημαίνει ότι το SVM είναι ευαίσθητο στον θόρυβο και στις ακραίες τιμές που είναι κοντά στο όριο απόφασης. Επίσης, ο αυξανόμενος αριθμός των ακραίων τιμών, οδηγεί σε σημαντική αύξηση του αριθμού των διανυσμάτων υποστήριξης. Για την αντιμετώπιση του προβλήματος των ακραίων τιμών και των αδυναμιών των SVMs αναπτύχθηκαν οι Ισχυρές Μέθοδοι Διανυσμάτων Υποστήριξης (Robust SVM - RSVM). Τα RSVM χρησιμοποιούν μια νέα με-

θοδολογία στην ταξινόμηση μιας κατηγορίας με την χρήση μεθόδων ισχυρών διανυσμάτων υποστήριξης. Η προτεινόμενη μεθοδολογία λαμβάνει υπόψη τον μέσο όρο των δειγμάτων που βοηθούν στη μείωση των ακραίων τιμών και συνεπώς στη μείωση του αριθμού των διανυσμάτων υποστήριξης. Επίσης μπορούν να χρησιμοποιηθούν για την δημιουργία ενός ισχυρού διαγράμματος ελέγχου που προσαρμόζεται στα πραγματικά δεδομένα και δεν εξαρτάται από τη κατανομή. Ένα τέτοιο διάγραμμα ελέγχου υπολογίζει ένα όριο ελέγχου που προσδιορίζει τις συνθήκες εκτός ελέγχου και καθορίζεται από την ισχυρή απόσταση.

Το RSVM διατηρεί την κύρια ιδέα ενός τυπικού SVM και χρησιμοποιεί το προσαρμοστικό περιθώριο για κάθε σημείο αναφοράς για να διατυπώσει το πρόβλημα ελαχιστοποίησης. Στο RSVM, το προσαρμοστικό περιθώριο διαμορφώνεται χρησιμοποιώντας την απόσταση μεταξύ του κέντρου κάθε τάξης των δεδομένων εκπαίδευσης και του σημείου του δείγματος. Η βέλτιστη συνάρτηση περιλαμβάνει μια νέα μεταβλητή slack που είναι το γινόμενο της προεπιλεγμένης παραμέτρου και της τετραγωνικής απόστασης μεταξύ κάθε σημείου αναφοράς και του κέντρου της αντίστοιχης κλάσης. Επιπλέον, οι Ισχυρές Μέθοδοι Διανυσμάτων Υποστήριξης χρησιμοποιούνται και για την επίλυση προβλημάτων over-fitting όπου οι ακραίες τιμές καθιστούν τις τάξεις μη διαχωρίσιμες. Στη μέθοδο RSVM, ελαχιστοποιείται



**Σχήμα 4.3:** Πρόβλημα over-fitting

μόνο το περιθώριο του βάρους και όχι το άθροισμα του περιθωρίου. Η ελαχιστοποιημένη αντικειμενική συνάρτηση δίνεται ως εξής:

$$\phi(\sigma) = \frac{1}{2}\sigma^T\sigma \tag{4.22}$$

με

$$\beta_i f(\alpha_i) \geq 1 - \psi\mu^2 \tag{4.23}$$

όπου

$$\mu^2 = |\phi(\alpha_i) - \phi(\alpha_i^{'})| \tag{4.24}$$

με $\alpha_1$ η αναπαράσταση δείγματος κλάσης 1, $\beta_1$ η αναπαράσταση δείγματος κλάσης 2, $\sigma$ το διάνυσμα βάρους και η Εξίσωση (4.23) το σφάλμα από το υπερεπίπεδο.

Η Εξίσωση (4.23) απεικονίζει το πρόβλημα ελαχιστοποίησης του περιθωρίου του βάρους ως εναλλακτική της ελαχιστοποίησης του αθροίσματος των σφάλμα περιθωρίου και του σφάλματος ταξινομήσης στο τυπικό SVM. Η νέα μεταβλητή slack που προστίθεται είναι η $\psi\mu^2$ με $\psi \geq 0$. Η προεπιλεγμένη αυτή παράμετρος αντιπροσωπεύει την κανονικοποιημένη απόσταση μεταξύ κάθε σημείου αναφοράς και του κέντρου της αντίστοιχης κλάσης. Τα $[\phi(\alpha_i)]_{i=1}^m$ είναι ένα σύνολο μη γραμμικών μετασχηματισμών από το χώρο εισόδου στο χώρο των χαρακτηριστικών (Vapnik 1998). Ένας χώρος δεδομένων μετατρέπεται σε ένα χώρο χαρακτηριστικών που έχει την ίδια διάσταση με αυτόν του αρχικού χώρου δεδομένων.

Η μέγιστη απόσταση μεταξύ του κέντρου και των δεδομένων εκπαίδευσης της αντίστοιχης τάξης δίνεται από το $\eta$. Όταν τα κανονικοποιημένα δεδομένα βρίσκονται στην επιφάνεια μιας υπερσφαίρας, τότε η απόσταση $\eta$ ικανοποιεί την ακόλουθη ανισότητα:

$$0 \leq \eta^2 \leq 1 \tag{4.25}$$

$$\beta_i f(\alpha_i) = 1 - \psi\mu^2 \tag{4.26}$$

Τα σημεία που ικανοποιούν τα παραπάνω είναι τα ισχυρά διανύσματα υποστήριξης.

Όσον αφορά το πρόβλημα βελτιστοποίησης, κατασκευάζουμε την ακόλουθη συνάρτηση απόφασης:

$$f(\alpha_i) = \sigma\phi(\alpha_i) + \delta \tag{4.27}$$

με $\delta$ η μεροληψία.

Για να απαλειφθεί η πολυπλοκότητα της μεθόδου ελαχιστοποίησης και για να μειωθεί ο αριθμός των μεταβλητών, το πρόβλημα ελαχιστοποίησης μετατρέπεται σε επίλυση ενός δυϊκού προβλήματος χρησιμοποιώντας τη μέθοδο των πολλαπλασιαστών Lagrange. Η Λαγκραντζιανή είναι:

$$\mathbf{J}(\sigma, \delta, \rho) = \frac{1}{2}\sigma^T\sigma - \sum_i^m \rho_i\big(\beta_i(\sigma\phi_(\alpha_i) + \delta) - 1 + \psi\mu^2\big) \tag{4.28}$$

όπου $\rho_i$ οι πολλαπλασιαστές Lagrange.

Ελαχιστοποιούμε την συνάρτηση Λαγκράντζ ως προς $\sigma$ και $\delta$ ως εξής:

$$\frac{\partial \mathbf{J}(\sigma, \delta, \rho)}{\partial \sigma} = 0 \tag{4.29}$$

$$\Rightarrow \sigma - \sum_{i}^{m} \rho_i \beta_i \phi(\alpha_i) = 0 \tag{4.30}$$

$$\Rightarrow \sigma = \sum_{i}^{m} \rho_i \beta_i \phi(\alpha_i) \tag{4.31}$$

και

$$\frac{\partial \mathbf{J}(\sigma, \delta, \rho)}{\partial \delta} = 0 \tag{4.32}$$

$$\Rightarrow - \sum_{i}^{m} \rho_i \beta_i = 0 \tag{4.33}$$

$$\Rightarrow \sum_{i}^{m} \rho_i \beta_i = 0 \tag{4.34}$$

Χρησιμοποιώντας αυτά τα αποτελέσματα στην Λαγκραντζιανή καταλήγουμε στην:

$$\mathbf{J}(\sigma, \delta, \rho) = \sum_{i}^{m} \rho_i (1 - \psi\mu^2) - \frac{1}{2} \rho_i \rho_j \beta_i \beta_j K(\alpha_i, \alpha_j) \tag{4.35}$$

Το δυϊκό πρόβλημα μετασχηματίζεται ως εξής:

$$\mathbf{J}(\rho) = \sum_{i}^{m} \rho_i (1 - \psi\mu^2) - \frac{1}{2} \rho_i \rho_j \beta_i \beta_j K(\alpha_i, \alpha_j) \tag{4.36}$$

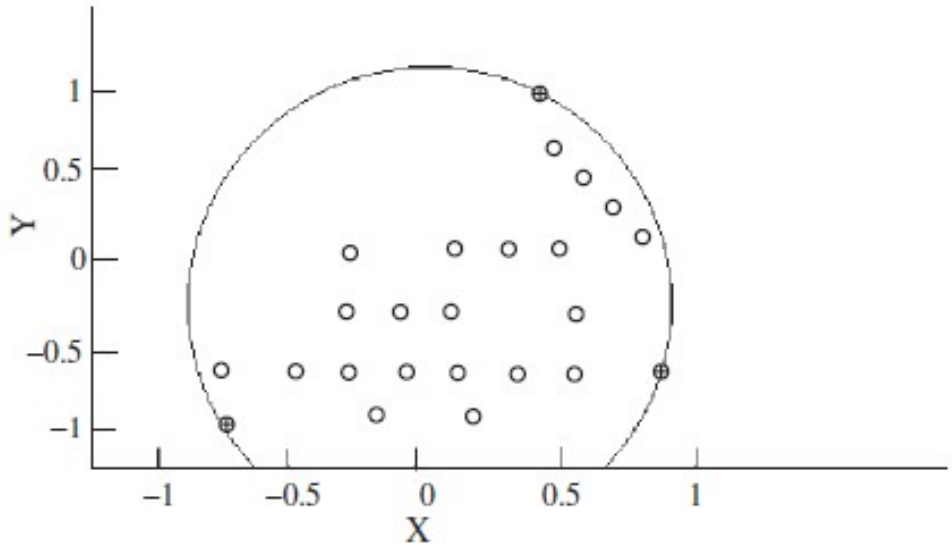
όπου

$$\sum_{i}^{m} \rho_i \beta_i = 0. \tag{4.37}$$

με $K$ η συνάρτηση πυρήνα.

Τα ισχυρά διανύσματα υποστήριξης χρησιμοποιούνται στην ταξινόμηση μιας τάξης με στόχο την εκτίμηση των πολυμεταβλητών κατανομών. Αυτά τα προβλήματα συναντώνται συχνά στη βιομηχανία και είναι αρκετά περίπλοκα λόγω των δυσκολιών που σχετίζονται με την εκτίμηση των πολλαπλών μεταβλητών. Για παράδειγμα, μια υπερσφαίραπερικλείει αρχικά τα πρωταρχικά δείγματα που συλλέκτηκαν κάτω από την κανονική κατανομή. Η υπερσφαίρα που κατασκευάζεται δεν αντικατοπτρίζει τη στενή περιγραφή του ορίου των

δειγμάτων (Σχήμα (4.4)). Για αυτόν τον λόγο επιλέγεται ένα όριο που βασίζεται στα ισχυρά διανύσματα υποστήριξης καθώς είναι πιο ευέλικτο και περικλείει τα δεδομένα έχοντας όσο το δυνατόν μικρότερο όγκο (Σχήμα (4.5)).

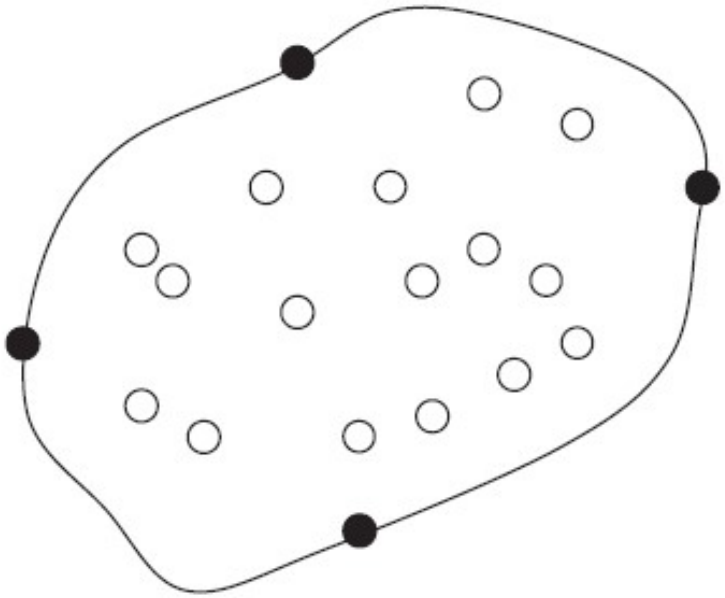



**Σχήμα 4.4:** Αναπαράσταση Υπερσφαίρας

Για την κατασκευή του προσαρμοστικού περιθωρίου, υπολογίζεται η απόσταση μεταξύ του κέντρου κάθε τάξης των δεδομένων εκπαίδευσης και του δείγματος. Στη βέλτιστη συνάρτηση, εισάγεται μια νέα μεταβλητή slack που είναι το γινόμενο της προεπιλεγμένης παραμέτρου $\psi$ καθώς και η τετραγωνική απόσταση μεταξύ κάθε σημείου αναφοράς και του κέντρου της αντίστοιχης τάξης. Επιπλέον, για την επίλυση του προβλήματος βελτιστοποίησης χρησιμοποιείται ένα σύνολο εκπαίδευσης που αφορά τον θόριβο. Αυτό καθιστά το προσαρμοστικό περιθώριο λιγότερο ευαίσθητο σε ακραίες τιμές.

Η επιλογή του ισχυρού διανύσματος υποστήριξης από όλα τα πρωταρχικά δείγματα γίνεται με τετραγωνικό προγραμματισμό και βασίζεται στο δυϊκό πρόβλημα που αναλύσαμε παραπάνω. Τα ισχυρά διανύσματα είναι αυτά τα σημεία του δείγματος που αντιστοιχούν στα θετικά $\rho_i$ και αυτά τα διανύσματα είναι πιο σημαντικά για τον προσδιορισμό του ορίου ελέγχου. Πιο συγκεκριμένα:

$$MAX! \quad \sum_{i}^{m} \rho_i(1 - \psi\mu^2) - \frac{1}{2}\rho_i\rho_j\beta_i\beta_j K(\alpha_i, \alpha_j) \tag{4.38}$$



**Σχήμα 4.5:** Όριο ελέγχου με RSVM

όπου

$$0 \leq \rho_i \leq \xi \tag{4.39}$$

και

$$\sum_{i}^{m} \rho_i = 1 \tag{4.40}$$

Το άνω όριο των παραμέτρων $\rho_i$ ελέγχεται από μια θετική σταθερά $\xi$. Το σφάλμα τύπου I θα είναι μεγαλύτερο στην περίπτωση που το $\xi$ είναι μικρό. Αυτό συμβαίνει γιατί πολλά $\rho_i$ βρίσκονται κοντά στο μηδέν. Στην περίπτωση που το $\xi$ είναι μεγαλύτερο, το σφάλμα τύπου II θα είναι επίσης μεγάλο λόγω της ευαισθησίας του ορίου. Η κατάλληλη τιμή του $\xi$ εξαρτάται κάθε φορά από την περίπτωση που βρισκόμαστε και η βέλτιστη τιμή του επιτυγχάνεται μετά από εκτεταμένους πειραματισμούς.

Η μεταβλητή slack $\psi\mu^2$ λαμβάνεται υπόψη ως μέρος του προσαρμοστικού περιθωρίου. Θεωρούμε ένα ακραίο σημείο. Σε σύγκριση με άλλα κανονικά σημεία στη δεδομένη τάξη, παρατηρείται ότι η απόσταση μεταξύ αυτού του σημείου και του κέντρου της τάξης είναι αρκετά μεγάλη. Η τιμή του όρου $\psi\mu^2$ είναι επίσης ελαφρώς μεγάλη. Έτσι, οι συντελεστές $\rho_i$ που σχετίζονται με το σημείο αναφοράς που ικανοποιεί την Ανισότητα (4.39) κινείται προς το μηδέν. Ως εκ τούτου, αυτό το σημείο αναφοράς δεν είναι διάνυσμα υποστήριξης. Με αυτόν τον τρόπο μειώνεται ο αριθμός των διανυσμάτων υποστήριξης και η απόκλιση του

προσαρμοστικού περιθωρίου.

Στην συνέχεια, αναφέρουμε την κατασκευή ενός ισχυρού διαγράμματος για τον προσδιορισμό του ορίου ελέγχου και την ανίχνευση των συνθηκών εκτός ελέγχου. Ο αρχικός στόχος είναι να εντοπιστούν τα δείγματα που είναι ισχυρα διανύσματα στήριξης. Το όριο ελέγχου δεν είναι ένα συνηθισμένο όριο και δεν μπορεί να εκφραστεί μαθηματικά. Λαμβάνοντας υπόψιν ότι η κατανομή δεν είναι κανονική ή δεν υπάρχει γνώση σχετικά με την κατανομή, εφαρμόζονται οι μέθοδοι ισχυρών διανυσμάτων υποστήριξης για τον προσδιορισμό του ορίου ελέγχου. Τα ισχυρά διανύσματα υπολογίζουν την απόσταση μεταξύ κάθε σημείου αναφοράς και του κέντρου της αντίστοιχης τάξης στον χώρο του πυρήνα και έτσι καθορίζεται το όριο ελέγχου. Η απόσταση, έστω $\mu$, υπολογίζεται ως εξής:

$$\phi(\alpha'_{\beta_i}) = \frac{1}{n}\sum_{l}\phi(\alpha_i), \tag{4.41}$$

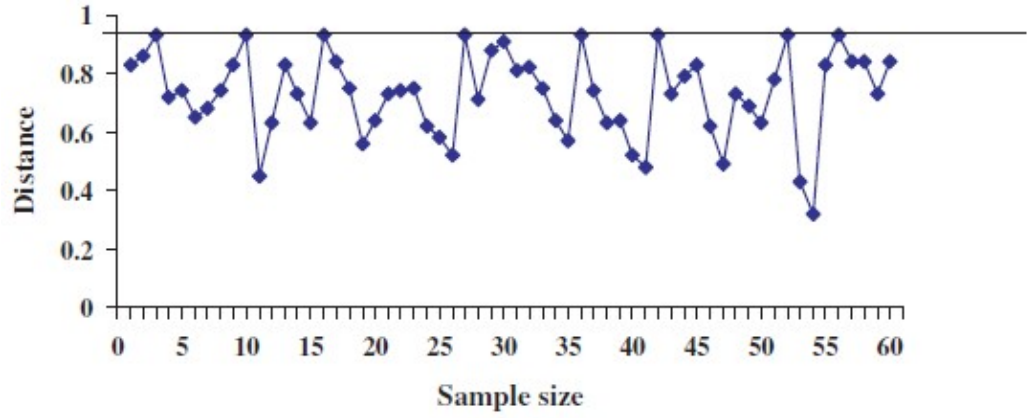
όπου $n$ ο αριθμός των σημείων δεδομένων, $\alpha'_{\beta_i}$ το κέντρο της κλάσης και

$$\mu = \frac{\sqrt{\nu.\nu - 2K(\nu, \alpha'_{\beta_i}) + K(\nu, \alpha'_{\beta_i}, \alpha'_{\beta_i})}}{\eta}. \tag{4.42}$$

Ο ρόλος μιας ταξινόμησης μιας κλάσης είναι να προσαρμόσει την απόκλιση και να βρει το καλύτερο όριο για τα δεδομένα. Σε ένα ισχυρό γράφημα, το περίγραμμα αντιπροσωπεύεται από μια οριζόντια γραμμή που αντιστοιχεί στην απόσταση του πυρήνα των ορίων και ως εκ τούτου το διάγραμμα μπορεί να εφαρμοστεί για την παρακολούθηση μιας διεργασίας και τον εντοπισμό των σημείων εκτός ελέγχου ανεξάρτητα από την κατανομή και το μέγεθος του δείγματος.

### 4.2.6 Διάγραμμα Ελέγχου Ανάλυσης Κύριας Συνιστώσας Πυρήνα

Ένα ακόμη διάγραμμα που εντάσεται στην κατηγορία των διαγραμμάτων πυρήνα είναι το Διάγραμμα Ελέγχου Ανάλυσης Κύριας Συνιστώσας του Πυρήνα (Kernel Principal Component Analysis Control Chart - KPCA). Το διάγραμμα αυτό βασίζεται στον υπολογισμό




**Σχήμα 4.6:** Μορφή ισχυρού διαγράμματος

του στατιστικού του Hotelling $T^2$ για τον πίνακα των κύριων συνιστωσών του πυρήνα, ο οποίος υπολογίζεται με την εφαρμογή του αλγόριθμου KPCA στη διασπορά και στον πίνακα συνδιακύμανσης για την παρατήρηση χαρακτηριστικών και ιδιοτήτων καθώς και για την ανακάλυψη εξωτερικών παραγόντων που επηρεάζουν τα όρια του διαγράμματος.

Αρχικά, θα πρέπει να παρουσιάσουμε το Διάγραμμα Ελέγχου Hotelling $T^2$ το οποίο αναπτύχθηκε για την ταυτόχρονη παρακολούθηση των μεταβλητών ποιότητας, διαστάσης $p$, μιας διαδικασίας πολλαπλών μεταβλητών. Το διάγραμμα προκύπτει με την προσαρμογή του στατιστικού $T^2$, ενά μέτρο απόστασης που βασίζεται στην κανονική κατανομή, στο γράφημα. Αν $X_1, X_2, ..., X_p$ είναι $p$ συσχετισμένα ποιοτικά χαρακτηριστικά και οι παραμέτροι είναι άγνωστοι, τότε το διάγραμμα διαμορφώνεται από τα ιστορικά δείγματα δεδομένων και το μέγεθος του δείγματος είναι 1. Το στατιστικό $T^2$ σε αυτή την περίπτωση δίνεται από τον τύπο:

$$T^2 = (X_i - \bar{X})'(S)^{-1}(X_i - \bar{X}), \tag{4.43}$$

όπου $S$ ο πίνακας διασποράς-συνδιακύμανσης του δείγματος και $\bar{X}$ ο διανυσματικός μέσος του δείγματος.

Τα άνω και κάτω όρια ελέγχου είναι αντίστοιχα:

$$UCL = \frac{(m-1)^2}{m}\beta_{\alpha,p/2,(m-p-1)/2} \tag{4.44}$$

$$LCL = 0 \tag{4.45}$$

όπου $m$ είναι το μέγεθος των παρατηρήσεων στα ιστορικά δεδομένα, $\beta_{\alpha,p/2,(m-p-1)/2}$ είναι η Βήτα κατανομή με παραμέτρους $p/2$ και $(m-p-1)/2$.

Για την υλοποίηση του διαγράμματος KPCA επιλέγουμε πρώτα τη συνάρτηση πυρήνα, η οποία εξαρτάται από την παράμετρο εξομάλυνσης $h$ ή το εύρος ζώνης, και στη συνέχεια υπολογίζουμε τον πίνακα του πυρήνα σύμφωνα με τον ακόλουθο γενικό τύπο:

$$K = K_{i,j} = \Omega(x_i)\Omega(x_j) \qquad i = 1, ..., n \qquad j = 1, ..., n. \tag{4.46}$$

Αφού προσδιοριστούν τόσο η συνάρτηση του πυρήνα όσο και ο πίνακας του πυρήνα, υπολογίζονται τα στοιχεία του πυρήνα από τον ακόλουθο τύπο:

$$p_\omega = \sum_{j=1}^{n} \alpha_{i,j} K(x_i, x_j) \tag{4.47}$$

και από την πρώτη κύρια συνιστώσα του $p$, το στατιστικό $T^2$ υπολογίζεται σύμφωνα με τον εξής τύπο:

$$\tilde{T}_K^2 = \sum_{i=1}^{l} p_i \lambda_i^{-1} p_i^T \tag{4.48}$$

όπου $\lambda_i$ η $i$-οστή χαρακτηριστική ρίζα.

Τα όρια ελέγχου για αυτό το διάγραμμα καθορίζονται από τον εκτιμητή πυκνότητας του πυρήνα, καθώς η κατανομή του στατιστικού $\tilde{T}_K^2$ είναι άγνωστη και μπορεί να εκφραστεί από την εξής εξίσωση:

$$\hat{f}_h = \frac{1}{m\hat{h}} \sum_{i=1}^{n} K \frac{T^2 - \tilde{T}_K^2}{\hat{h}}, \tag{4.49}$$

όπου $K$ η συνάρτηση πυρήνα και $\hat{h}$ η παράμετρος εξομάλυνσης. Επιπλέον, η συνάρτηση κατανομής της $\hat{f}_h$ μπορεί να υπολογιστεί από την:

$$\hat{F}_h(\tilde{T}_K^2) = \int_0^{\tilde{t}_K^2} \hat{f}_h(\tilde{T}_K^2) d(\tilde{T}_K^2). \tag{4.50}$$

Γνωρίζουμε ότι το $\hat{F}_h(\tilde{T}_K^2)$ υπολογίζεται σύμφωνα με τον κανόνα του τραπεζίου, ο οποίος είναι μία από τις αριθμητικές μεθόδους ολοκλήρωσης, εξαρτάται από τη διαμερισμό της περιόδου ολοκλήρωσης σε σημεία που έχουν καθοριστεί εκ των προτέρων με σκοπό τη

λήψη κατά προσέγγιση μαθηματικών τύπων για τον υπολογισμό των ολοκληρωμάτων ως εξής:

$$\int_{p_{min}}^{p_{max}} \hat{f}_h(\tilde{T}_K^2)d(\tilde{T}_K^2) \approx \frac{\pi_{max} - \pi_{min}}{2n} \sum_{i=1}^{n} (\hat{f}_h(\tilde{T}_{K,i}^2) + \hat{f}_h(\tilde{T}_{K,i+1}^2)), \tag{4.51}$$

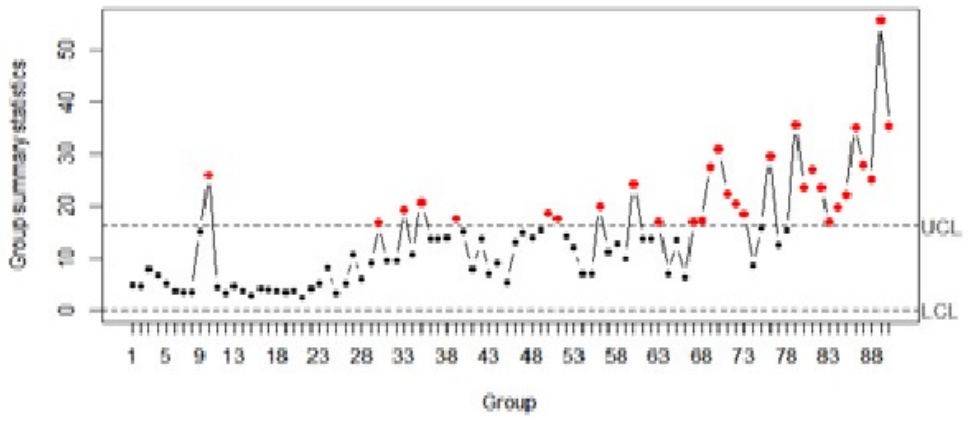
όπου $\pi_{max}, \pi_{min}$ αντιπροσωπεύουν τη μεγαλύτερη και τη μικρότερη τιμή του $\tilde{T}_K^2$, επομένως το ανώ όριο ελέγχου προσδιορίζεται με τον υπολογισμό της ποσοστιαίας τιμής $(100(1-\alpha)^{th})$ και ο μέσος όρου ψευδούς συναγερμού $\alpha$, του οποίου η τιμή κυμαίνεται μεταξύ $1 < \alpha < 0$ υπολογίζεται σύμφωνα με τον ακόλουθο τύπο:

$$CL = \hat{F}_h(\tilde{t}_k^2)^{-1}(1-\alpha). \tag{4.52}$$

Η συνάρτηση πυρήνα που επιλέγεται είναι αυτή του $Gauss$ με τύπο:

$$k(x_i, x_j) = exp\bigg(-\frac{\|x_i - x_j\|^2}{2h^2}\bigg) \tag{4.53}$$

Στην συνέχεια θα παρουσιάσουμε τον αλγόριθμο KPCA σε έξι βήματα:




**Σχήμα 4.7:** Παράδειγμα Διαγράμματος KPCA

1. Επιλέγουμε την συνάρτηση πυρήνα και υπολογίζουμε τον πίνακα του πυρήνα.
2. Υπολογίζουμε την κύρια συνιστώσα του πυρήνα από την Εξίσωση (4.47).
3. Υπολογίζουμε το χαρακτηριστικό $\tilde{T}_K^2$ των πρώτων $l$ κύριων συνιστωσών $p$ από την Εξίσωση (4.48).

4. Υπολογίζουμε το $\hat{f}_h$ το οποίο είναι ο εκτιμητής της πυκνότητας για το $\tilde{T}_K^2$.

5. Υπολογίζουμε το $\hat{F}_h(\tilde{T}_K^2)$ και προσδιορίζουμε το όριο ελέγχου από την Εξίσωση (4.52).

6. Σχεδιάζουμε το διάγραμμα με βάση το $\tilde{T}_K^2$ και την κεντρική γραμμή (central line - CL). Το CL είναι το άνω όριο ελέγχου και η διαδικασία θεωρείται εκτός ελέγχου αν ξεπεράσει μια στατιστική τιμή του $\tilde{T}_K^2$.

**Πίνακας 4.1:** Σύγκριση ARL των διαγραμμάτων KPCA και KNN

| Alpha | KPCA | KNN |
|---|---|---|
| 0.5 | 1.486 | 1.611 |
| 0.1 | 8.194 | 8.975 |
| 0.05 | 18.4 | 18.317 |
| 0.01 | 98.889 | 95.556 |
| 0.005 | 197.778 | 197.778 |

### 4.2.7 Τυχαίο KNN Διάγραμμα

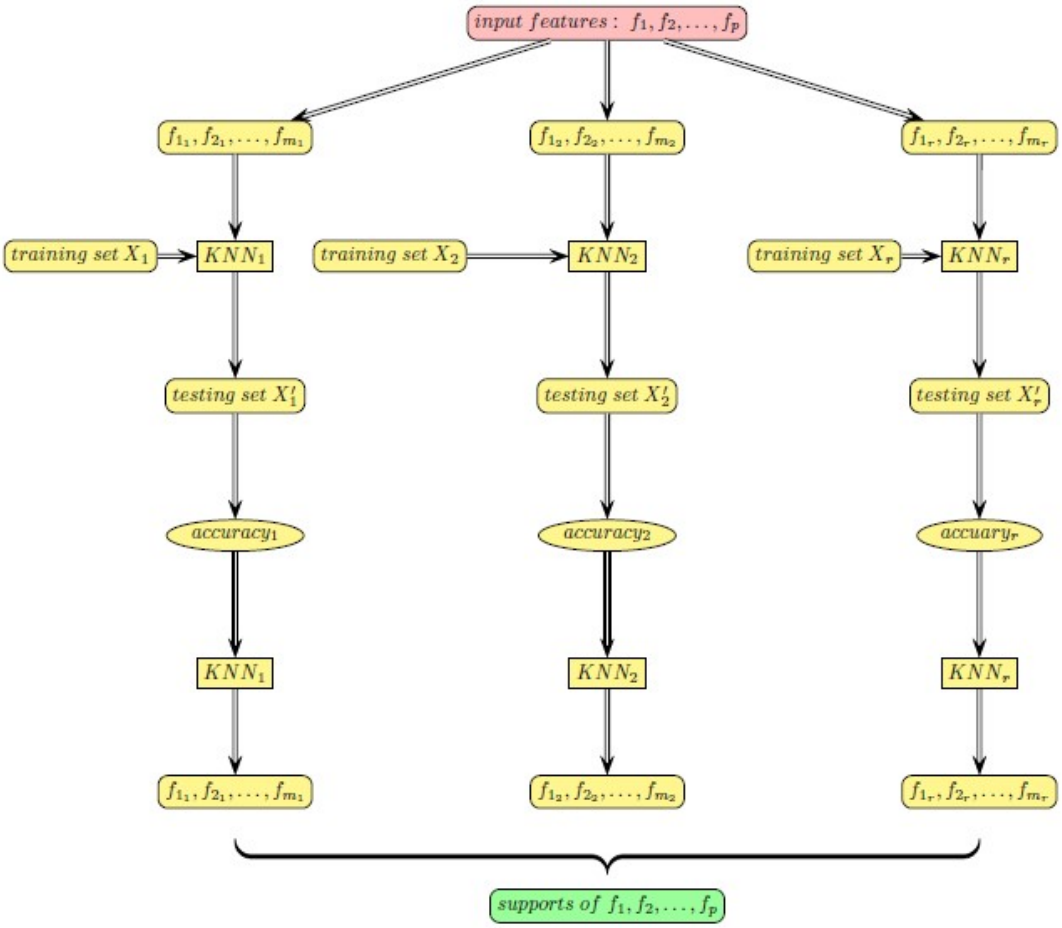
Το Τυχαίο KNN Διάγραμμα (Random KNN - RKNN ) αποτελεί μια γενίκευση του αλγορίθμου του $k$-πλησιέστερου γείτονα KNN. Είναι ειδικά σχεδιασμένο για ταξινόμηση σε σύνολα δεδομένων υψηλών διαστάσεων. Το RKNN περιλαμβάνει πολλά από τα πλεονεκτήματα του KNN. Συγκεκριμένα, το KNN είναι μια μη παραμετρική μέθοδος ταξινόμησης. Δεν λαμβάνει καμία παραμετρική μορφή για την κατανομή των τυχαίων μεταβλητών. Λόγω της ευελιξίας του μη παραμετρικού μοντέλου, είναι συνήθως ένας καλός ταξινομητής για πολλές καταστάσεις στις οποίες η κοινή κατανομή είναι άγνωστη ή είναι δύσκολο να μοντελοποιηθεί παραμετρικά. Αυτό ισχύει ιδιαίτερα για σύνολα δεδομένων υψηλών διαστάσεων. Ένα άλλο σημαντικό πλεονέκτημα του KNN είναι ότι οι τιμές που λείπουν (missing values) μπορούν εύκολα να υπολογιστούν. Επιπλέον, το KNN είναι γενικά πιο ισχυρό και πιο ευαίσθητο σε σύγκριση με άλλους δημοφιλείς ταξινομητές. Με βάση τα παραπάνω πλεονεκτήματα, το RKNN οδηγεί σε σημαντική βελτίωση της απόδοσης όσον αφορά την υπολογιστική πολυπλοκότητα, αλλά και την ακρίβεια της ταξινόμησης.

Η ιδέα του RKNN βασίζεται στην τεχνική των RF και είναι παρόμοια με τη μέθοδο της τυχαίας επιλογής υποχώρου που χρησιμοποιείται για τα Δάση Απόφασης (Decision

Forests). Τόσο τα RF όσο και τα Δάση Απόφασης χρησιμοποιούν δέντρα απόφασης ως βασικούς ταξινομητές. Το RKNN χρησιμοποιεί το KNN ως βασικό ταξινομητή, χωρίς να εμπλέκεται η ιεραρχική δομή. Σε σύγκριση με τα δέντρα αποφάσεων, το KNN είναι απλό στην εφαρμογή και είναι ιδιαίτερα σταθερό. Το RKNN μπορεί να σταθεροποιηθεί με έναν μικρό αριθμό βασικών KNN και ως εκ τούτου θα χρειαστεί στην συνέχεια μόνο ένας μικρός αριθμός σημαντικών μεταβλητών. Αυτό σημαίνει ότι το τελικό μοντέλο θα είναι απλούστερο συγκριτικά με τα RF ή τα Δάση Απόφασης. Συγκεκριμένα, θα δημιουργηθεί μια συλλογή από $r$ διαφορετικούς ταξινομητές KNN. Κάθε μία παίρνει ένα τυχαίο υποσύνολο των μεταβλητών εισόδου. Δεδομένου ότι το KNN είναι σταθερό, δεν κρίνεται απαραίτητη η χρήση μιας bootstrap διαδικασίας. Κάθε ταξινομητής KNN ταξινομεί ένα σημείο δοκιμής με βάση την πλειοψηφία ή την τάξη σταθμισμένης πλειοψηφίας των $k$ πλησιέστερων γειτόνων του. Η τελική ταξινόμηση σε κάθε περίπτωση καθορίζεται από την πλειοψηφία των ταξινομητών $r$. Αυτό μπορεί να αναφερθεί και ως η ψήθος της πλειοψηφίας μιας πλειοψηφίας (Majority of a Majority). Δηλαδή, έστω $\mathbf{F} = f_1, f_{2,...,fp}$ τα $p$ χαρακτηριστικά εισόδου, και $\mathbf{X}$ τα $n$ αρχικά διανύσματα δεδομένων εισόδου μήκους $p$, δηλαδή ένας πίνακας $n \times p$. Για δοσμένο $m < p$, συμβολίζουμε με $\mathbf{F}^{(m)} = \{f_{j1}, f_{j2}, ..., fjm | f_{jl} \in \mathbf{F}, 1 \leq l \leq m\}$ ένα τυχαίο υποσύνολο που προέρχεται από το $\mathbf{F}$ ισοπίθανα. Ομοίως, έστω $\mathbf{X}^{(m)}$ τα διανύσματα δεδομένων στον υποχώρο που όριζεται από το $\mathbf{F}^{(m)}$, δηλαδή ένας πίνακας $n \times m$. Τότε ένας ταξινομητής $KNN^{(m)}$ κατασκευάζεται εφαρμόζοντας τον βασικό αλγόριθμο KNN σε μια τυχαία συλλογή χαρακτηριστικών στο $\mathbf{X}^{(m)}$. Μία συλλογή από $r$ τέτοιους ταξινομητές συνδυάζεται για να δημιουργηθεί ο τελικός ταξινομητής RKNN.

Για να επιλέξουμε ένα υποσύνολο μεταβλητών προς ταξινόμηση, είναι απαραίτητο να καθοριστούν ορισμένα κριτήρια για την κατάταξη των μεταβλητών. Ορίζουμε ένα μέτρο, που ονομάζεται υποστήριξη (support). Κάθε χαρακτηριστικό $f$ θα εμφανίζεται σε ορισμένους ταξινομητές KNN, έστω στο σετ $\mathbf{C}(f)$ μεγέθους $M$, όπου $M$ είναι η πολλαπλότητα του $f$. Στην συνέχεια, κάθε ταξινομητής $c \in \mathbf{C}(f)$ είναι ένας αξιολογητής των $m$ χαρακτηριστικών του, έστω το σύνολο $\mathbf{F}(c)$. Μπορούμε να λάβουμε την ακρίβειά του ως μέτρο απόδοσης για αυτά τα χαρακτηριστικά. Η μέση ακρίβεια αυτών των ταξινομητών KNN (υποστήριξη) είναι ένα μέτρο της σύδεσης των χαρακτηριστικών με το αποτέλεσμα. Έτσι έχουμε μια κατάταξη των χαρακτηριστικών. Αυτή την διαδικασία την ονομάζουμε αμφίδρομη ψηφοφορία (bidirectional voting). Κάθε χαρακτηριστικό συμμετέχει τυχαία σε μια σειρά από διάφορα

KNN για να δώσει την ψήφο του για την ταξινόμηση. Με τη σειρά του, κάθε αποτέλεσμα ταξινόμησης ψηφίζει για κάθε χαρακτηριστικό που συμμετέχει.



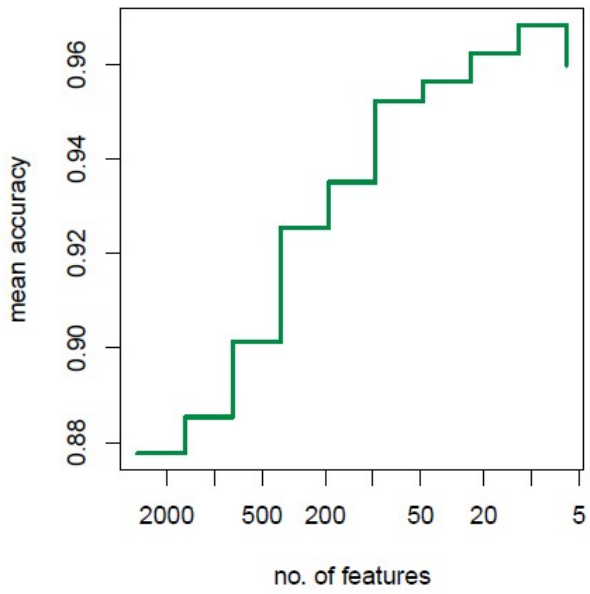

**Σχήμα 4.8:** Αμφίδρομη ψηφοφορία σε RKNN

Για τον υπολογισμό των υποστηρικτικών χαρακτηριστικών, τα δεδομένα χωρίζονται σε υποσύνολα βάσης και ερωτημάτων (query). Μπορούν να χρησιμοποιηθούν δύο μέθοδοι διαχωρισμού:

1. δυναμικός διαχωρισμός: Για κάθε KNN, οι περιπτώσεις διαμερίζονται τυχαία. Το ένα μισό είναι το βασικό υποσύνολο και το άλλο μισό είναι το υποσύνολο ερωτημάτων.

2. το σύνολο δεδομένων διαμερίζεται μία φορά, και για όλα τα KNN, χρησιμοποιούνται τα ίδια υποσύνολα βάσης και ερωτημάτων. Δηλαδή, όλα τα βασικά υποσύνολα και όλα τα υποσύνολα ερωτημάτων είναι ίδια.

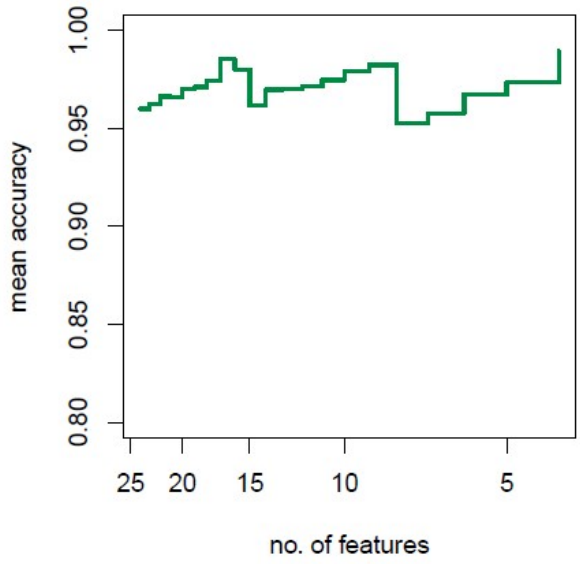
Για λόγους ποικιλομορφίας των KNN, προτιμάται ο δυναμικός διαχωρισμός.




**Σχήμα 4.9:** Διακύμανση μέσης ακρίβειας 1ο Στάδιο

Με τις υποστηρίξεις των χαρακτηριστικών, μπορούμε να επιλέξουμε απευθείας χαρακτηριστικά υψηλής κατάταξης αφού εκτελέσουμε τον αλγόριθμο υποστήριξης σε ολόκληρο το σύνολο δεδομένων. Αυτή την ονομάζουμε άμεση επιλογή (direct selection). Αλλά αυτή η απλή προσέγγιση μπορεί να επιφέρει πολλά ρίσκα για δεδομένα υψηλών διαστάσεων. Ακολουθούμε μια πιο συντηρητική και ασφαλέστερη προσέγγιση και εφαρμόζουμε αναδρομικά την άμεση διαδικασία επιλογής. Για να ισορροπήσουμε την σχέση μεταξύ ταχύτητας και απόδοσης ταξινόμησης, χωρίζουμε την αναδρομή σε δύο στάδια. Το πρώτο στάδιο είναι γρήγορο και ο αριθμός των μεταβλητών μειώνεται κατά μια δεδομένη αναλογία (1/2 από προεπιλογή). Αυτό το στάδιο είναι μια διαδικασία γεωμετρικής εξάλειψης (geometric elimination) αφού η διάσταση που πρέπει να διατηρηθεί είναι μια γεωμετρική πρόοδος. Στο δεύτερο στάδιο, ένας σταθερός αριθμός χαρακτηριστικών (1 από προεπιλογή) απορρίπτεται κάθε φορά. Αυτή είναι μια διαδικασία γραμμικής αναγωγής. Τέλος, ένα σχετικά μικρό σύνολο μεταβλητών θα επιλεγεί για τα τελικά μοντέλα. Επιπλέον, απαιτείται ένα άλλο κριτήριο αξιολόγησης για το σύνολο των χαρακτηριστικών. Χρησιμοποιούμε τη μέση ακρίβεια των $r$ τυχαίων KNN. Μετά το πρώτο στάδιο, μπορούμε να σχεδιάσουμε τη μέση ακρίβεια σε σχέση με τον αριθμό των χαρακτηριστικών. Η προσέγγιση λίγο πριν επιτευχθεί η μέγιστη ακρίβεια ονομάζεται προ-μέγιστη προσέγγιση (pre-max iteration). Το σύνολο των χαρακτηριστικών από την προ-μέγιστη προσέγγιση θα είναι η είσοδος για την επιλογή του δεύτερου σταδίου. Το RKNN έχει τρεις παραμέτρους, τον αριθμό των πλησιέστερων γειτόνων $k$, τον αριθμό

των τυχαίων KNN $r$ και τον αριθμό των χαρακτηριστικών για κάθε βασικό KNN $m$. Για τα σύνολα δεδομένων με μικρό $n$, και μεγάλα $p$, το $k$ θα πρέπει να είναι μικρό, π.χ. 1 ή 3, καθώς οι ομοιότητες μεταξύ των σημείων δεδομένων σχετίζονται με το πόσο κοντά βρίσκονται. Για τον $m$, προτείνεται να ισχύει $m = \sqrt{p}$ για να μεγιστοποιείται έτσι η διαφορά μεταξύ των υποσυνόλων των χαρακτηριστικών. Η απόδοση γενικά βελτιώνεται με την αύξηση του $r$, ωστόσο, πέρα από ένα σημείο, μεγαλύτερες τιμές του $r$ ενδέχεται να μην οδηγήσουν σε περαιτέρω βελτιώσεις. Πέρα από $r > 1000$, δεν υπάρχει κάποιο κέρδος όσον αφορά την ακρίβεια ταξινόμησης.




**Σχήμα 4.10:** Διακύμανση μέσης ακρίβειας 2ο Στάδιο

Οι Li, Harner Adjeroh (2011) εφάρμοσαν το παραπάνω διάγραμμα στα σύνολα δεδομένων λευχαιμίας του Golub. Το Σχήμα (4.9) δείχνει τη διακύμανση της μέσης ακρίβειας με μειωμένο αριθμό χαρακτηριστικών στο πρώτο στάδιο της επιλογής χαρακτηριστικών. Το Σχήμα (4.10) δείχνει τη διακύμανση της μέσης ακρίβειας με μειωμένο αριθμό χαρακτηριστικών στο δεύτερο στάδιο. Στο Σχήμα (4.10), η μέγιστη μέση ακρίβεια επιτυγχάνεται όταν υπάρχουν 4 γονίδια στο μοντέλο. Αυτά τα τέσσερα τελευταία γονίδια που επιλέχθηκαν για ταξινόμηση λευχαιμίας είναι: X95735at, U27460at, M27891at L09209sat. Χρησιμοποιώντας αυτά τα τέσσερα γονίδια και τον ταξινομητή KNN ($\kappa = 3$) για την ταξινόμηση των 34 ανεξάρτητων δειγμάτων δοκιμής, 18 από τις 20 περιπτώσεις ταξινομούνται σωστά. Η συνολική ακρίβεια είναι 91%. Αυτό το μοντέλο είναι πολύ απλό σε σύγκριση με άλλα που χρησιμοποιούν πολύ περισσότερα γονίδια.

# Κεφάλαιο 5

# Παράδειγμα Πραγματικών Δεδομένων

Στο κεφάλαιο αυτό, θα παρουσιάσουμε μια εφαρμογή των διαγραμμάτων K-chart, KNN chart και KM-chart, τα οποία έχουν αναλυθεί στο Κεφάλαιο 5. Η εφαρμογή τους γίνεται πάνω σε ένα σύνολο πραγματικών δεδομένων από το εργοστάσιο καπνού Kairouan Tobacco Manufacture. Αφού γίνει αναλυτική παρουσίαση του συνόλου δεδομένων, θα αναλυθούν τα διαγράμματα και τα αποτελέσματά τους και τέλος θα υπάρξη σύγκριση στην αποτελεσματικότητα και ακρίβειά τους.

## 5.1 Σύνολο Δεδομένων

Η Kairouan Tobacco Manufacture (KTM) και η National Tobacco and Matches Corporation αντιπροσωπεύουν το μονοπώλιο του καπνού στην Τυνησία. Η KTM εξειδικεύεται στην παραγωγή ποικιλίας καπνού, κυρίως τσιγάρων, πούρων, καπνού για πίπες και καπνού με ταμπάκο, σύμφωνα με τα εθνικά πρότυπα. Σε αυτήν την εφαρμογή, μας ενδιαφέρει η ποιότητα των τσιγάρων «Cristal Light», που αρχικά μελετήθηκαν από τον Hajlaoui (2011). Με στόχο τη βελτίωση της ποιότητας των προϊόντων, την ικανοποίηση των καταναλωτών και τη συμμόρφωση με την ισχύουσα νομοθεσία, η KTM εξόπλισε τα εργαστήριά της με εξοπλισμό που επιτρέπει τον αυστηρό ποιοτικό έλεγχο σε όλες τις φάσεις της παραγωγικής

διαδικασίας. Πράγματι, περισσότερες από 16000 χημικές και φυσικές αναλύσεις πραγματοποιούνται κάθε χρόνο στα εργαστήρια της εταιρείας. Στην ενότητα, δίνουμε μια λεπτομερή περιγραφή της διαδικασίας παραγωγής των τσιγάρων Cristal Light, η οποία αποτελείται απο 12 στάδια:

1. Ύγρανση των φύλλων καπνού με στόχο τη διατήρηση του σωστού επιπέδου σχετικής υγρασίας περιβάλλοντος στο 60-70% και την ισορροπία μεταξύ του αέρα και της υγρασίας στον καπνό.

2. Αλώνισμα των φύλλων καπνού όπου η σκόνη και τα ξένα αντικείμενα αφαιρούνται από τα φύλλα, έτσι ώστε να παραμείνει μόνο ο καθαρός καπνός.

3. Επεξεργασία λωρίδων, η οποία στοχεύει στην επεξεργασία του καπνού σε φύλλα κόβοντάς τον σε λωρίδες καπνού. Ταξινομούνται σε βαρύτερο και ελαφρύτερο κλάσμα.

4. Κατακερματισμός με ειδική μηχανή για τον τεμαχισμό του καπνού.

5. Στέγνωμα, όπου ο καπνός που προέρχεται από τον τεμαχιστή ξηραίνεται σε ένα ειδικό μηχάνημα.

6. Επέκταση των άκρων. Σε αυτό το βήμα, μια μηχανή προσθέτει διογκωμένα στελέχη καπνού στον προηγουμένως κομμένο και αποξηραμένο καπνό.

7. Περίβλημα που αποτελείται από μια σάλτσα που χρησιμοποιείται για να αναδείξει τη γεύση του καπνού. Η σάλτσα είναι μείγμα νερού και μίας επιλογής φυσικών συστατικών. Ο σκοπός του περιβλήματος είναι διπλός: να αναδείξει τη γεύση και να διατηρήσει τον καπνό υγρό όσο το δυνατόν περισσότερο.

8. Αρωματισμός για να βελτιωθεί η γεύση του καπνού.

9. Εισαγωγή διογκωμένου καπνού, δηλαδή μετά την ολοκλήρωση της παρασκευής και επεξεργασίας του καπνού, είναι έτοιμος για τη διαδικασία παραγωγής.

10. Παρασκεύασμα τσιγάρων, όπου ο τεμαχισμένος καπνός διαμορφώνεται σε δύο μικρούς, συνεχείς, διπλούς κυλίνδρους κομμένου καπνού με χαρτί γύρω τους. Τα φίλτρα προστίθενται στον καπνό χρησιμοποιώντας ειδική μηχανή και έχουμε το τελικό τσιγάρο.

11. Συσκευασία και εγκιβωτισμός που γίνεται εντελώς αυτοματοποιημένα, με ένα μηχάνημα που τοποθετεί τα τσιγάρα στα πακέτα.

12. Τελική προετοιμασία όπου τα πακέτα των τσιγάρων τυλίγονται χρησιμοποιώντας μια ειδική μηχανή, προσθέτει επίσης την κρατική σφραγίδα και ένα μικρό λουράκι για το άνοιγμα της συσκευασίας.

Οι δραστηριότητες ποιοτικού ελέγχου καλύπτουν όλα τα στάδια της παραγωγικής διαδικασίας. Ξεκινούν με την επίβλεψη της κατασκευής φύλλων καπνού, μέχρι την προετοιμασία των τσιγάρων σε πακέτα. Στη βιομηχανία καπνού, η ποιότητα ορίζεται από πολλές απαιτήσεις που επιβάλλονται από το κράτος. Αυτές οι απαιτήσεις λαμβάνονται υπόψη κατά τον καθορισμό των προδιαγραφών τσιγάρων. Ως εκ τούτου, ορίζονται όρια ανοχής για κάθε μέτρο ποιοτικού χαρακτηριστικού των τσιγάρων. Η εταιρεία KTM χρησιμοποιεί πέντε χαρακτηριστικά που διασφαλίζουν την ποιότητα των τσιγάρων:

1. Το βάρος ενός τσιγάρου, το οποίο αποτελείται από τον καπνό, το φίλτρο και τα βάρη τσιγαρόχαρτου. Κυμαίνεται μεταξύ 0,965 και 1 γραμμαρίου.

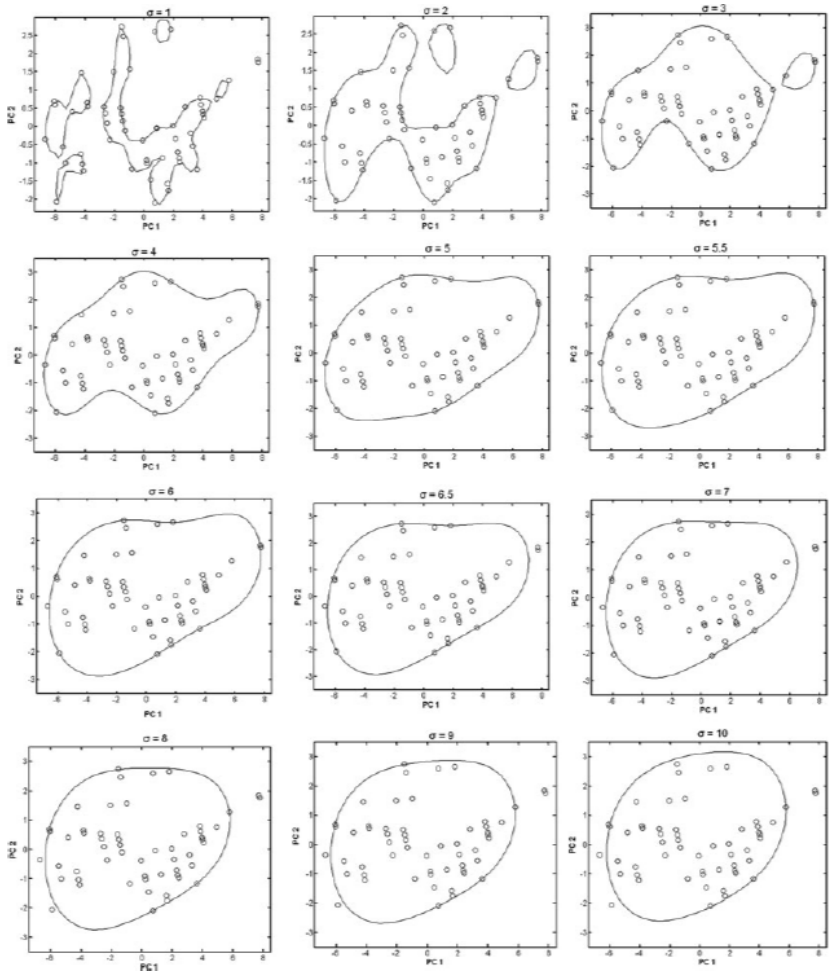
2. Το δομοστοιχείο ενός τσιγάρου, που αντιστοιχεί στη διάμετρό του. Κυμαίνεται από 6,75 έως 8,0 χιλιοστά.

3. Το ποσοστό υγρασίας του καπνού, το οποίο είναι η αναλογία νερού που περιέχεται σε ένα τσιγάρο. Θεωρείται αποδεκτό εάν κυμαίνεται μεταξύ 11,5% και 13,5%.

4. Αντίσταση έλξης ενός τσιγάρου, η οποία ορίζεται από τη διαφορά πίεσης μεταξύ των δύο άκρων ενός τσιγάρου όταν διέρχεται ποσότητα αέρα από αυτό. Η αντίσταση έλξης θεωρείται αποδεκτή όταν κυμαίνεται από 100 έως $115CE$

5. Πυκνότητα αναδίπλωσης, δηλαδή ο όγκος που καταλαμβάνει η μάζα του καπνού μέσα σε ένα τσιγάρο($450 \pm 20\ cm^3$)

Το σύνολο των δεδομένων αποτελείται από 65 παρατηρήσεις. Τα πρώτα 60 τσιγάρα χρησιμοποιούνται για την κατασκευή διαγραμμάτων μίας κλάσης στη φάση I. Κάθε τσιγάρο χρειάστηκε ένα λεπτό για να συλλεχθεί. Τα υπόλοιπα πέντε τσιγάρα χρησιμοποιούνται για τον έλεγχο καταστάσεων εκτός ελέγχου στη φάση II.




**Σχήμα 5.1:** SVDD κλάσεις για διάφορες τιμές του $\sigma$

## 5.2 Κατασκευή Διαγραμμάτων

Προκειμένου να κατασκευάσουμε και τα τρία διαγράμματα $K$,$KNN$ και $KM$, πρέπει να ακολουθήσουμε τα ίδια τρία βασικά βήματα. Αρχικά, το σύνολο των δεδομένων αναλύεται χρησιμοποιώντας τη μέθοδο ανάλυσης κύριας συνιστώσας (PCA) για να ληφθούν ανεξάρτητα και ομοίως κατανεμημένα δεδομένα, μια θεμελιώδης υπόθεση για το πρόβλημα ταξινόμησης μιας κατηγορίας. Τα κύρια στοιχεία (PCs) που προκύπτουν από το πρώτο βήμα χρησιμοποιούνται για την κατασκευή της καταληλότερης κλάσης. Στην εφαρμογή μας, έχουμε τρεις ταξινομητές μιας κατηγορίας, τους SVDD, KNNDD και KMDD. Η βέλτιστη κλάση που λαμβάνεται από το δεύτερο βήμα χρησιμοποιείται για την κατασκευή των διαγραμμάτων μιας κλάσης με τον υπολογισμό στατιστικών γραφημάτων. Όλοι οι υπολογισμοί έγιναν με το MATLAB. Λεπτομέρειες για τον κώδικα υπάρχουν στο Παράθεμα (5.2.4) από τους Gani και Limam M.(2012).

### 5.2.1 Διάγραμμα K-chart

Το $K$ διάγραμμα, όπως έχουμε αναφέρει, χρησιμοποιεί τον αλγόριθμο $SVDD$. Για να κατασκευάσουμε την βέλτιστη κλάση μίας κατηγορίας χρησιμοποιούμε την Γκαουσιανή συνάρτηση πυρήνα

$$K(x_i, x_j) = exp\bigg( - \frac{\|x_i - x_j\|^2}{\sigma^2} \bigg), \tag{5.1}$$

όπου $\sigma > 0$ είναι το πλάτος του πυρήνα που ελέγχει τα όρια του. Για να υπολογίσουμε την ιδανική τιμή του $\sigma$ χρησιμοποιούμε το κριτήριο $F_{measure}$ που ορίζεται ως εξής:

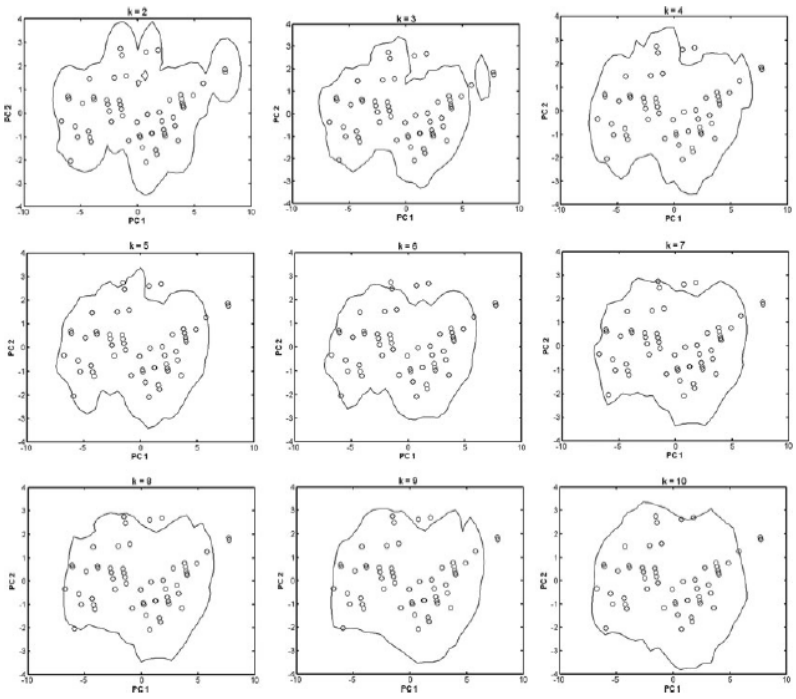
$$F_{measure} = \frac{2 \times precision \times recall}{precision + recall}, \tag{5.2}$$

όπου $0 \leq F_{measure} \leq 1$, $precision = \frac{truepositionrate}{truepositionrate + falsepositionrate}$ και $recall = 1 - precision$. Όσο μεγαλύτερη είναι η τιμή του $F_{measure}$, τόσο καλύτερο είναι το $\sigma$ που της αντιστοιχεί. Στην συνέχεια υπάρχει ο πίνακας με τους υπολογισμούς για το βέλτιστο $\sigma$.

**Πίνακας 5.1:** Χαρακτηριστικά SVDD ταξινομητών μιας κλάσης

| $\sigma$ | SV | $F_{measure}$ |
|---|---|---|
| 1.00 | 36 | 0.835 |
| 1.50 | 25 | 0.878 |
| 2.00 | 22 | 0.909 |
| 2.50 | 16 | 0.974 |
| 3.00 | 13 | 0.938 |
| 3.50 | 10 | 0.956 |
| 4.00 | 8 | 0.956 |
| 4.50 | 9 | 0.965 |
| 5.00 | 8 | 0.956 |
| 5.50 | 8 | 0.947 |
| 6.00 | 8 | 0.947 |
| 6.50 | 8 | 0.947 |
| 7.00 | 8 | 0.947 |
| 7.50 | 8 | 0.947 |
| 8.00 | 9 | 0.947 |
| 8.50 | 9 | 0.938 |
| 9.00 | 9 | 0.947 |
| 9.50 | 9 | 0.947 |
| 10.00 | 8 | 0.965 |

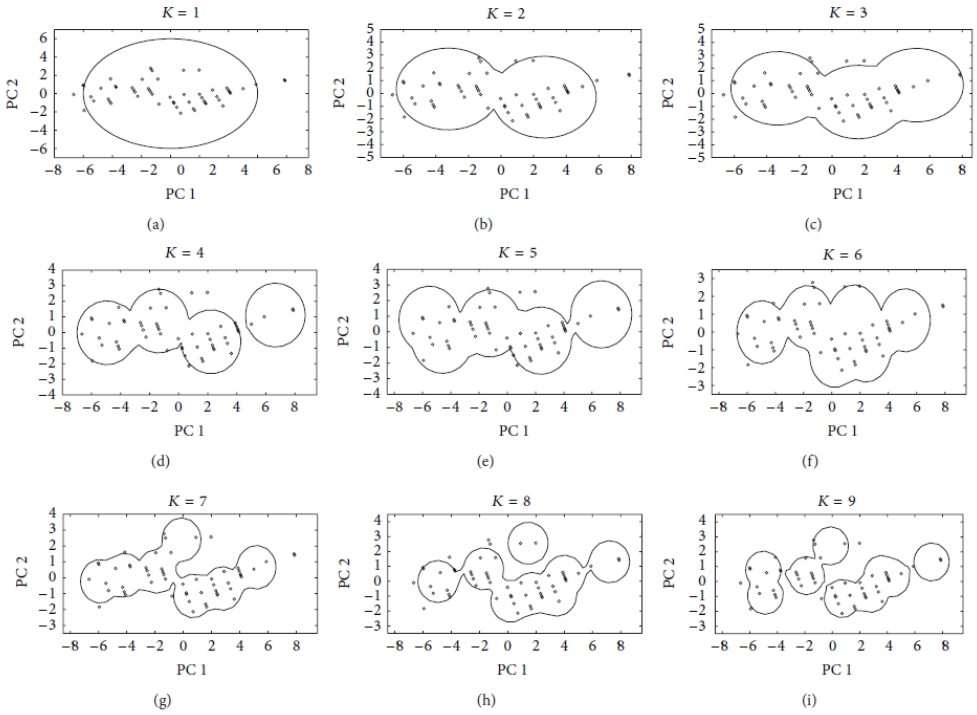
Η ευκρίνεια (precision) είναι μια μέτρηση που ποσοτικοποιεί τον αριθμό των σωστών θετικών προβλέψεων που έγιναν και υπολογίζει την ακρίβεια για την τάξη μειοψηφίας. Η




**Σχήμα 5.2:** KNN κλάσεις για διάφορες τιμές του $k$

ανάκληση (recall) είναι μια μέτρηση που ποσοτικοποιεί τον αριθμό των σωστών θετικών προβλέψεων που έγιναν από όλες τις θετικές προβλέψεις που θα μπορούσαν να είχαν γίνει. Σε αντίθεση με την ευκρίνεια που αφορά μόνο τις σωστές θετικές προβλέψεις από όλες τις θετικές προβλέψεις, η ανάκληση παρέχει μια ένδειξη χαμένων θετικών προβλέψεων. Το κριτήριο $F_{measure}$ παρέχει έναν τρόπο συνδυασμού της ακρίβειας και της ανάκλησης σε ένα ενιαίο μέτρο που καταγράφει και τις δύο ιδιότητες. Από μόνη της, ούτε η ευκρίνεια ούτε η ανάκληση παρέχουν την πλήρη εικόνα του προβλήματος. Το $F_{measure}$, που είναι ο αρμονικός μέσος, παρέχει έναν τρόπο έκφρασης και των δύο.

Η βέλτιστη τιμή του πλάτους του πυρήνα Gauss είναι $\sigma = 10$ καθώς αντιστοιχεί στην υψηλότερη τιμή του $F_{measure}$, ίση με 0.965, όπως φαίνεται από τον Πίνακα 5.1. Γενικά όμως, μόνο η τιμή του $F_{measure}$ δεν αρκεί για την επιλογή του βέλτιστου πυρήνα. Ένα άλλο κριτήριο είναι η γραφική αναπαράσταση της κλάσης. Είναι σαφές από το Σχήμα 5.1 ότι για $\sigma = 1, 2, 3$ η κατασκευή κλάσης δεν είναι δυνατή. Επιπλέον, παρατηρούμε ότι η ομαλότητα των κλάσεων ενισχύθηκε όταν αυξήθηκε η τιμή του $\sigma$.




**Σχήμα 5.3:** KMDD κλάσεις για διάφορες τιμές του $K$

### 5.2.2 Διάγραμμα KNN-chart

Για την κατασκευή της βέλτιστης μίας κλάσης που βασίζεται στον ταξινομητή KNNDD, θα πρέπει να καθοριστεί το βέλτιστο μέγεθος του πλησιέστερου γείτονα, που συμβολίζεται με $k$. Στην πραγματικότητα, η παράμετρος $k$ παίζει κεντρικό ρόλο στην απόδοση μίας κλάσης που βασίζεται στον KNNDD. Αυτό συμβαίνει διότι το $k$ καθορίζει την αντιστάθμιση μεταξύ της υπερομαλότητας και της υποομαλότητας του ορίου ελέγχου. Αυτό μπορεί να φανεί γραφικά, αν συγκρίνει κανείς το πρώτο με το τελευταίο γράφιμα στο Σχήμα 5.2

Είναι σαφές επίσης από το Σχήμα 5.2 ότι κάτω από διαφορετικές τιμές του $k$, λαμβάνονται διαφορετικά σχήματα της κλάσης. Το σωστό εύρος του $k$ είναι μεταξύ 10 και 50. Ένας άλλος παράγοντας που μπορεί να επηρεάσει την παράμετρο $k$ είναι το μέγεθος του δείγματος. Όσο μικρότερο είναι το μέγεθος των δεδομένων εκπαίδευσης, τόσο μικρότερο είναι το $k$. Στην εφαρμογή μας, η βέλτιστη κλάση λαμβάνεται θέτοντας $k = 10$.

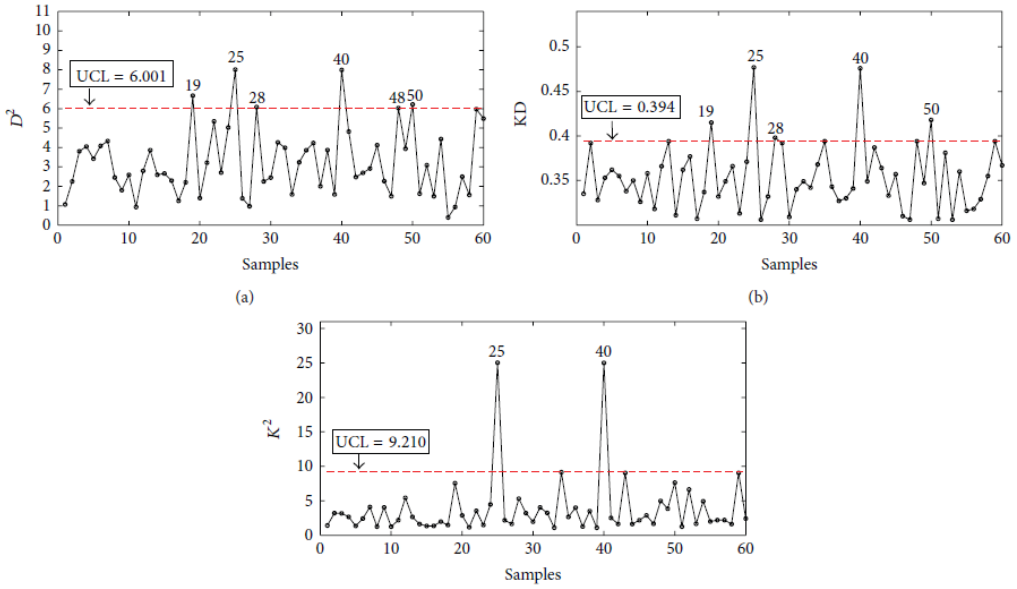
### 5.2.3 Διάγραμμα KM-chart

Μετά την εκτέλεση του PCA, δοκιμάστηκαν αρκετοί αριθμοί συστάδων για την κατασκευή της μιας κλάσης που βασίζεται στον ταξινομητή KMDD. Είναι σαφές από το Σχήμα 5.3 ότι ο αριθμός των συστάδων επηρεάζει το σχήμα της κλάσης και παίζει καθοριστικό ρόλο στον προσδιορισμό της αντιστάθμισης μεταξύ της ομαλότητας και της υποομαλότητας του ορίου ελέγχου. Στην περίπτωση μας, επιλέγουμε $K = 1$ καθώς το δείγμα δεν είναι υπερβολικά μεγάλο.

Η ανίχνευση μιας μη φυσιολογικής παρατήρησης στην τάξη-στόχο εξαρτάται από το σχήμα της τάξης. Το KMDD παρείχε μια σφαιρική κλάση, ενώ η SVDD έδωσε μια ευέλικτη μη σφαιρική κλάση λόγω της χρήσης διανυσμάτων υποστήριξης. Αξίζει να σημειωθεί ότι το σχήμα μιας κλάσης που βασίζεται σε SVDD εξαρτάται από το πλάτος της συνάρτησης πυρήνα, ενώ το σχήμα μιας κλάσης που βασίζεται σε KNNDD είναι συνάρτηση του μεγέθους του πλησιέστερου γείτονα.

### 5.2.4 Ανάλυση Αποτελεσμάτων

Στη Φάση I, χρησιμοποιείται ένα δείγμα 60 τσιγάρων Cristal Light για τη ρύθμιση της διαδικασίας ελέγχου. Τα τρία διγράμματα δίνουν διαφορετικά αποτέσματα, γεγονός αναμενόμενο, καθώς βασίζονται σε διαφορετικούς ταξινομητές. Το διάγραμμα K κατασκευάζεται με έξι διανύσματα υποστήριξης που δίνουν $UCL = 0,394$. Ανιχνεύει πέντε τσιγάρα εκτός ελέγχου, τα τσιγάρα 19, 25, 28, 40 και 50. Το διάγραμμα KNN είναι κατασκευασμένο με $UCL = 9.210$ και ανιχνεύει μόνο δύο τσιγάρα εκτός ελέγχου, συγκεκριμένα, τα τσιγάρα 40 και 50, όπως φαίνεται στο Σχήμα 5.4. Το διάγραμμα $KM$ ξεπέρασε το όριο ελέγχου $UCL = 6.001$ περίπου στο 19ο, 25ο, 28ο, 40ο, 48ο και 50ο τσιγάρο, όπως φαίνεται στο Σχήμα 2. Για αυτά τα εκτός ελέγχου τσιγάρα, τουλάχιστον ένα από τα πέντε ποιοτικά χαρακτηριστικά τους δεν τηρούσε κάποιο διάστημα ανοχής, όπως αναφέρουμε στην σελίδα 82. Σε σύγκριση με τα δύο άλλα διαγράμματα ελέγχου, το διάγραμμα KM κατάφερε να εντοπίσει μία νέα εκτός ελέγχου παρατήρηση, το τσιγάρο 48. Και το διάγραμμα $K$ και το διάγραμμα KNN απέτυχε να το ανιχνεύσει. Μόλις αφαιρεθούν αυτές οι εκτός ελέγχου παρατηρήσεις, δεν εντοπίζονται επιπλέον ακραίες τιμές και η διαδικασία είναι εντός ελέγχου.




**Σχήμα 5.4:** Φάση I

Στη Φάση II, χρησιμοποιήθηκαν πέντε τσιγάρα Cristal Light για τον εντοπισμό καταστάσεων εκτός ελέγχου. Το KM διάγραμμα εντόπισε το 62ο τσιγάρο εκτός ελέγχου. Το 62ο τσιγάρο ήταν εκτός ελέγχου και για το διάγραμμα K, ενώ τα τσιγάρα με αριθμό 62 και 65 είναι εκτός ελέγχου από το διάγραμμα KNN. Το Σχήμα 5.5 φαίνονται τα διαγράμματα για τη Φάση II.

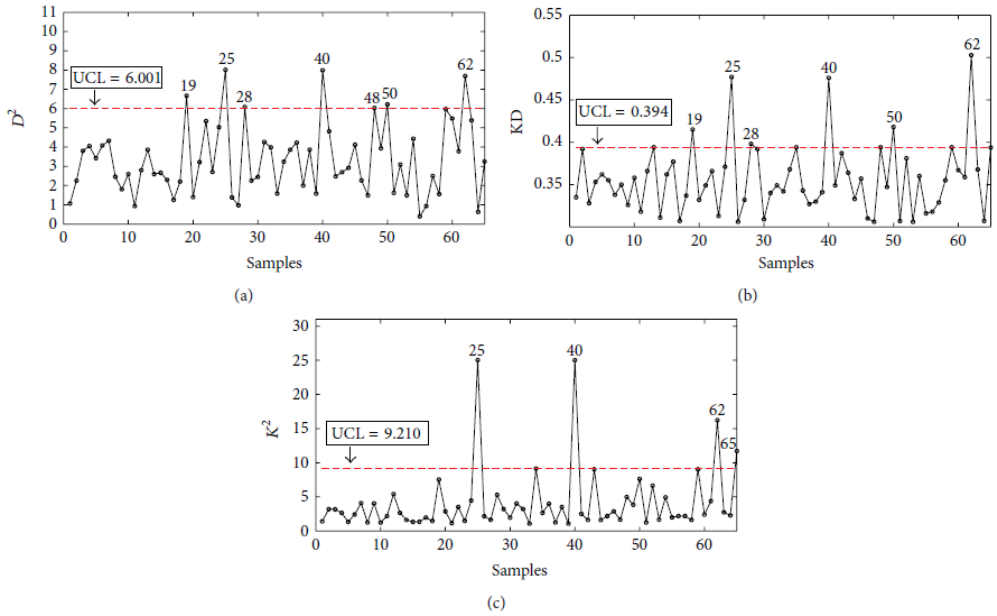
Στην συνέχεια, μελετάμε την απόδοση του διαγράμματος KM, K και KNN. Η μελέτη απόδοσης βασίζεται στο κριτήριο ARL. Δίνεται το

$$ARL = \frac{1}{p},$$

όπου $p$ είναι η πιθανότητα ένα σημείο να βρίσκεται εκτός ελέγχου.

Για την εκτίμηση του ARL, οι Gani και Limam M. (2014) διεξήγαγαν μια μελέτη προσομοίωσης.

1. Δημιουργούνται πέντε πολυμεταβλητές κανονικές μεταβλητές με ένα μέσο διάνυσμα $\mu$ και έναν πίνακα συνδιακύμανσης $\Sigma$, παρόμοιο με το μέσο διάνυσμα και τον πίνακα συνδιακύμανσης των δεδομένων εντός ελέγχου που χρησιμοποιούνται στη βιομηχανική εφαρμογή. Το διάγραμμα K, KM και KNN έχουν σχεδιαστεί για να επιτυγχάνουν




**Σχήμα 5.5:** Φάση II

ένα συνολικό ARLίσο με 200. Η τιμή ARL υπολογίζεται με τον μέσο όρο των μηκών εκτέλεσης που λαμβάνονται με την εκτέλεση 10000 προσομοιωμένων διαγραμμάτων.

2. Υπολογίζουμε τις πολυμεταβλητές μετατοπίσεις, που συμβολίζονται με $\delta$, που γίνονται στο διάνυσμα του μέσου, σύμφωνα με τον Πίνακα 5.2. Μεγάλες τιμές του $\delta$ αντιστοιχούν σε μεγάλες μετατοπίσεις του μέσου. Η τιμή $\delta = 0$ είναι η κατάσταση ελέγχου.

Για τον εντοπισμό μικρών μετατοπίσεων, π.χ. $\delta = 0.25$ το διάγραμμα KM έχει καλύτερη απόδοση από το γράφημα KNN, καθώς έδωσε ARL = 192.308, ενώ το γράφημα KNN έδωσε ARL = 200. Για την ίδια μετατόπιση, το διάγραμμα K έχει ARL = 100, καλύτερο από αυτό του διαγράμματος KM.

Για τον εντοπισμό μέτριων μετατοπίσεων, π.χ. $\delta = 1$ το διάγραμμα KNN αποδίδει καλύτερα συγκριτικά με τα άλλα δύο διαγράμματα. Έδωσε ARL = 40, ενώ οι τιμές των ARL του διαγράμματος K και KM αντίστοιχα ήταν 50 και 147.059.

Η διαφορά στην ευαισθησία στις μετατοπίσεις στον μέσο σε κάθε διάνυσμα οφείλεται στον τρόπο που ορίζεται η απόσταση σε καθένα από τους ταξινομητές τους. Στο διάγραμμα K χρησιμοποιείται ο KD, ενώ στο διάγραμμα KM και στο KNN χρησιμοποιείται η Ευκλείδεια

**Πίνακας 5.2:** Σύγκριση ARL

| $\delta$ | K-chart | KNN-chart | KM-chart |
|---|---|---|---|
| 0.00 | 200.000 | 200.000 | 200.000 |
| 0.25 | 100.000 | 200.000 | 192.308 |
| 0.50 | 66.667 | 200.000 | 185.185 |
| 0.75 | 66.667 | 100.000 | 166.667 |
| 1.00 | 50.000 | 40.000 | 147.059 |
| 1.25 | 50.000 | 14.285 | 117.647 |
| 1.50 | 40.000 | 9.091 | 86.956 |
| 1.75 | 28.571 | 3.703 | 60.976 |
| 2.00 | 22.222 | 1.695 | 39.370 |
| 2.25 | 13.333 | 1.020 | 25.707 |
| 2.50 | 6.896 | 1.000 | 16.340 |
| 2.75 | 4.167 | 1.000 | 9.597 |
| 3.00 | 2.439 | 1.000 | 5.540 |
| 3.25 | 1.709 | 1.000 | 3.250 |
| 3.50 | 1.197 | 1.000 | 1.997 |
| 3.75 | 1.036 | 1.000 | 1.390 |
| 4.00 | 1.000 | 1.000 | 1.116 |
| 5.00 | 1.000 | 1.000 | 1.000 |

απόσταση. Το πλεονέκτημα της KD σε σχέση με την Ευκλείδεια απόσταση οφείλεται στην συνάρτηση πυρήνα. Αυτή, στο διάγραμμα K ισούται με απόσταση μεταξύ δύο δειγμάτων μετρημένα σε χώρο υψηλότερων διαστάσεων. Αυτό επιτρέπει το διάγραμμα K να ανιχνεύει ευκολότερα οποιαδήποτε μικρή αλλαγή στον μέσο. Επιπλέον, τα διανύσματα υποστήριξης παίζουν σημαντικό ρόλο στον καθορισμού του ορίου της κλάσης και είναι ευαίσθητα σε μικρές μετατοπίσεις. Σε ότι αφορά το ARL, ακόμα και για μικρές μετατοπίσεις στον μέσο, είναι εμφανές από τον Πίνακα 5.2 πως οι τιμές του ARL για το διάγραμμα KM βρίσκονται μεταξύ των τιμών του ARL για τα διαγράμματα K και KNN, με μεγαλύτερη τιμή να έχει πάντα το ARL του K διαγράμματος.

Εν κατακλείδι, σε γενικές γραμμές κάθε διάγραμμα έχει τα πλεονεκτήματά και τα μειονεκτήματά του. Για παράδειγμα, το K διάγραμμα έχει καλύτερα αποτέσματα από το διάγραμμα KM και KNN στον γρήγορο εντοπισμό μετατοπίσεων στη διαδικασία, ενώ το υπολογιστικό κόστος του διαγράμματος KM και του KNN είναι χαμηλότερο από αυτό του διαγράμματος K.

# Παράρτημα

## Κώδικας Διαγράμματος ΚΜ

%Let Q be the (n..×..m) matrix of quality variables where n is the number

%of observations and m the number of quality variables in the training phase.

%Define the target class.

$T = target\_class(+Q);$

%Use the KMDD algorithm to fit a sphere around the defined target class above,

%where c1 is a fraction error on the target class and c2 is a parameter

%defining the number of clusters.

$w = kmeans\_dd(T, c1, c2);$

%Show the results of KMDD classifier.

$W = +w;$

%Phase I of the KM-chart:

%Compute the Euclidean distance of each training observation and the UCL.

$n = size(T, 1);$

$D\_training = sqrt(min(sqeucldistm(+T, W.w), , 2))repmat(W.threshold, n, 1);$

%Phase II of the KM-chart:

%Let now R be the (k x p) matrix of quality variables where k is the number

%of observations and p the number of quality variables in the testing phase.

%In Phase II we repeat the same computation as in Phase I but here we use

%test data and compare it with the UCL to detect out-of-control states.

%Compute the Euclidean distance of each test observation.

$m = size(R, 1);$

$D\_test = sqrt(min(sqeucldistm(+R, W.w), , 2))repmat(W.threshold, m, 1);$

$y1 = D\_training(:, 1); y2 = D\_training(:, 2); x1 = (1 : n)';$

$y3 = D\_training(:, 1); D\_test(:, 1);$

$y4 = D\_training(:, 2); D\_test(:, 2); x2 = (1 : n + m)';$

%Display the KM-chart for Phase I and II.

figure;

$SUBPLOT(2, 1, 1), plot(x1, y1, `-o', x1, y2, `-'); title(`PhaseIoftheKM - chart')$

$SUBPLOT(2, 1, 2), plot(x2, y3, `-o', x2, y4, `-'); title(`PhaseIIoftheKM - chart')$

# Κώδικας Διαγράμματος K

%Let Q be the (nxm) matrix of quality variables where n is the number

%of observations and m the number of quality variables in the training phase

%define the target class

$T = target\_class(+Q);$

%use the SVDD method to fit a sphere around the defined target class above,

%where c1 is a fraction error on the target class and c2 is a parameter

%defining the width of the RBF kernel function

$w = svdd(T, c1, c2);$

% show the results of SVDD classifier

$W = +w;$

%phase I of the K chart:

% compute the matrix of KDs

$n = size(T, 1);$

$K\_training = exp(-sqeucldistm(+T, W.sv)/(W.s * W.s));$

% show the KD of each training observation

$KD\_training = W * offs\_2 * sum(repmat(W.a0, n, 1)\_K\_training, 2);$

% show the KD of each training observation and the UCL.

$KD\_training\_UCL = [KD\_trainingrepmat(W.threshold, n, 1)];$

%Phase II of the K chart:

%Let now R be the (k _p) matrix of quality variables where k is the number

%of observations and p the number of quality variables in the testing phase

%In phase II, we repeat the same computation as in phase I but here we will

%use test data and compare it with the UCL to detect out-of-control states

%compute the matrix of KDs for the test observations

$m1 = size(R, 1)$

$K\_test = exp(-sqeucldistm(+R, W.sv)/(W.s * W.s))$

%show the KD of each test observation

$KD\_test = W.offs\_2 * sum(repmat(W.a0, m1, 1)\_K\_test, 2)$

%show the KD of each test observation and the UCL

$KD\_test\_UCL = [KD\_testrepmat(W.threshold, m1, 1)]$

$y1 = KD\_training\_UCL(:, 1); y2 = KD\_training\_UCL(:, 2); x1 = (1 : n)'$

$y3 = [KD\_training\_UCL(:, 1); KD\_test\_UCL(:, 1)]$

$y4 = [KD\_training\_UCL(:, 2); KD\_test\_UCL(:, 2)]; x2 = (1 : n + m1)'$

%Display the K chart for phases I and II

figure;

$SUBPLOT(2, 1, 1), plot(x1, y1, '\_o', x1, y2, '\_'); title('PhaseIoftheKchart')$

$SUBPLOT(2, 1, 2), plot(x2, y3, '\_o', x2, y4, '\_'); title('PhaseIIoftheKchart')$

# Κώδικας Διαγράμματος KNN

%Let Q be the (n X m) matrix of quality variables where n is the number

%of observations and m the number of quality variables in the training

%phase.

%define the target class.

$T = target\_class(+Q);$

%use the KNNDD method to fit a sphere around the defined target class

%above, where c1 is a fraction error on the target class, c2 is

%the size of nearest neighbour and 'kappa' is a distance measure.

$w = knndd(T, c1, c2, 'kappa');$

%show the results of KNND classifier.

$W = +w;$

%Phase I of the KNN chart:

% compute the Euclidean distances of training observations.

$[n, m] = size(T);$

$distmat = sqeucldistm(+T, W.x);$

%show the matrix of the computed distances.

$[sD, I] = sort(distmat, 2);$

%use 'kappa' measure to compute the distance to the kth nearest neighbour.

$ind = sD(:, W.k);$

%compute the K-square statistics and the UCL.

$Ksquare\_training = [indrepmat(W.threshold, [n, 1])];$

%Phase II of the KNN chart:

%Let now R be the (n1 X m1) matrix of quality variables where n1 is the number

%of observations and m1 the number of quality variables in the testing phase.

%In Phase II we repeat the same computation as in Phase I but here we will

%use test data and compare it with the UCL to detect out-of-control states.

%compute the Euclidean distances of test observations.

$[n1, m1] = size(R);$

$distmat = sqeucldistm(+R, W.x);$

%show the matrix of the computed distances.

$[sD, I] = sort(distmat, 2);$

%use 'kappa' measure to compute the distance to the kth nearest neighbour.

$ind = sD(:, W.k);$

%compute the K-square statistics.

$Ksquare\_test = [indrepmat(W.threshold, [n1, 1])];$

$y1 = Ksquare\_training(:, 1); y2 = Ksquare\_training(:, 2); x1 = (1 : n)';$

$y3 = [Ksquare\_training(:, 1); Ksquare\_test(:, 1)];$

$y4 = [Ksquare\_training(:, 2); Ksquare\_test(:, 2)]; x2 = (1 : n + n1)';$

%Display of the K chart for phases I and II.

figure;

$SUBPLOT(2, 1, 1), plot(x1, y1, ' - o', x1, y2, '-'); title('PhaseIoftheKNNchart')$

$SUBPLOT(2, 1, 2), plot(x2, y3, ' - o', x2, y4, '-'); title('PhaseIIoftheKNNchart')$

# Βιβλιογραφία


[1] Adhikari R., Agrawal R. (2013). An Introductory Study on Time series Modeling and Forecasting. *arXiv preprint arXiv:1302.6613., 1–68.*

[2] Alfaro-Cortés E., Alfaro-Navarro J. L., Gámez M., García N. (2020). Using Random Forest to Interpret Out-of-Control Signals. *Acta Polytechnica Hungarica*. 17(6): 115-130.

[3] Chen S., Yu J. (2019). Deep recurrent neural network-based residual control chart for autocorrelated processes. *Qual Reliab Engng Int.* 35: 2687– 2708.

[4] Demircioglu D. D., Boran S., Cil I. (2020). Integration of machine learning techniques and control charts in multivariate processes. *Scientia Iranica*. 27(6): 3233-3241.

[5] Gani W., Limam M. (2013). Performance Evaluation of One-Class Classification-based Control Charts through an Industrial Application. *Quality and Reliability Engineering International*. 29(6), 841-854.

[6] Gani W., Limam M. (2014). A One-Class Classification-Based Control Chart Using the K-Means Data Description Algorithm *Journal of Quality and Reliability Engineering*. 2014

[7] Hajlaoui M. (2011). On the Charting Procedures: $T^2$ Chart and DD-Diagram. *Hindawi Publishing Corporation, International Journal of Quality, Statistics, and Reliability*. doi:10.1155/2011/830764.

[8] Hu J., Runger G. (2010). Time-based detection of changes to multivariate patterns. *Annals of Operations Research*. 174: 67-81

[9] Kuter, S., Usul, N., Kuter, N., (2011). Bandwidth Determination for Kernel Density Analysis of Wildfire Events at Forest Sub-District Scale. *Ecological Modelling*. 222(17): 3033-3040.

[10] Lee S., Lee S., Kim C.K. (2022). One-class classification-based monitoring for the mean and variance of time series. *Qual Reliab Eng Int.* 38(5), 2548-2565

[11] Montgomery D. C. (2013). *Introduction to Statistical Quality Control. 7th ed.* Hoboken: J.Wiley & Sons.

[12] Murakami Y., Mizuguchi K. (2010). Applying the Naive Bayes Classifier with Kernel Density Estimation to the Prediction of Protein-Protein Interaction Sites. *Bioinformatics*, 26(15): 1841-1848 .

[13] Neuhardt J. B. (1987). Effects of Correlated Sub-Samples in Statistical Process Control. *IIE Transactions*. 19(2): 208-214.

[14] Li, S., Harner, E.J., Adjeroh, D.A. (2011). Random KNN feature selection - a fast and stable alternative to Random Forests. *BMC Bioinformatics* 12, 450.

[15] Sun R., Tsung F. (2003). A Kernel-Distance-Based Multivariate Control Chart Using Support Vector Methods. *International Journal of Production Research*. 41(13):2975-2989

[16] Tran, P.H., Ahmadi Nadi, A., Nguyen, T.H., Tran, K.D., Tran, K.P. (2022). Application of Machine Learning in Statistical Process Control Charts: A Survey and Perspective. In: *Tran, K.P. (eds) Control Charts and Machine Learning for Anomaly Detection in Manufacturing*. Springer Series in Reliability Engineering. Springer.

[17] Vapnik V. (1998), *Statistical Learning Theory*. John Wiley  Sons, Chichester.

[18] Viharos Z. J., Jakab R. (2021). Reinforcement Learning for Statistical Process Contrôlin Manufacturing. *Measurement*. 182:109616

[19] Wang D. (2015) *Time Series Analysis*. Lecture Notes. Department of Statistics University of South Carolina.

[20] Weese M., Martinez W., Megahed F. M., Jones-Farmer L. A. (2016). Statistical Learning Methods Applied to Process Monitoring: An Overview and Perspective. *Journal of Quality Technology.* 48(1): 4-24.

[21] Woodward W. A., Gray H. L., Elliott A. C. (2017). *Applied time series analysis with R*, Second edition, Boca Raton: Taylor Francis, CRC Press.

[22] Xie X., Qui P. (2022) Machine Learning Control Charts for Monitoring Serially Correlated Data, In: *Tran K.P. (eds) Control Charts and Machine Learning for Anomaly Detection in Manufacturing (pp. 131-147).* Springer Series in Reliability Engineering. Springer, Cham.

[23] Yang K., Hancock W. M. (1990). Statistical quality control for correlated samples. *International Journal of Production Research.* 28(3): 595-608.