ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

Διπλωματική Εργασία

Μηχανές διανυσμάτων υποστήριξης και Στατιστικός έλεγχος διεργασιών

Ελένη Δίλμα
Α.Μ. : 09111026

Επιβλέπων: Κουκουβίνος Χρήστος,
Καθηγητής Ε.Μ.Π.

Αθήνα, 2017

National Technical University of Athens

School of Applied Mathematical and Physical Sciences

Thesis

Support Vector Machines and Statistical Process Control

Eleni Dilma

Supervisor: Koukouvinos Christos
Professor N.T.U.A.

Athens, 2017

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον καθηγητή του Εθνικού Μετσόβιου Πολυτεχνείου, κ. Χρήστο Κουκουβίνο, για την ανάθεση της συγκεκριμένης ενδιαφέρουσας εργασίας, την καθοδήγησης και τις συμβουλές του.

Παράλληλα, ιδιαίτερες ευχαριστίες θα ήθελα να εκφράσω στην υποψήφια διδάκτορα Αγγελική Λάππα για την ουσιαστική και πολύτιμη βοήθεια που μου παρείχε καθ΄όλη τη διάρκεια εκπόνησης της εργασίας μου.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου για τη διαρκή στήριξη και κατανόηση που μου έδειξε όλα τα χρόνια των σπουδών μου, καθώς και τον Πάνο, χωρίς τη βοήθεια και τη συμπαράσταση του οποίου η συγγραφή της συγκεκριμένης εργασίας θα ήταν πολύ δύσκολη.

Περίληψη

Ο στατιστικός έλεγχος διεργασιών είναι μία από τις μεγαλύτερες τεχνολογικές ανακαλύψεις του 20ου αιώνα και βοηθάει στην επίτευξη της ακρίβειας σταθερότητας της διεργασίας και στη βελτίωση της ικανότητας της μειώνοντας τη μεταβλητότητα. Ο ΣΕΔ ανιχνεύει έγκαιρα την εμφάνιση των αιτιών της μεταβλητότητας, ώστε να γίνουν οι διορθωτικές ενέργειες, προτού κατασκευαστούν προϊόντα που δεν πληρούν τις προδιαγραφές. Το πιο χρήσιμο εργαλείο είναι το διάγραμμα ελέγχου, το οποίο βελτιώνει την παραγωγικότητα, προβλέπει τα ελαττώματα, αποτρέπει την άσκοπη προσαρμογή και παρέχει πληροφορίες σχετικά με την ικανότητα της διεργασίας.

Οι μηχανές διανυσμάτων υποστήριξης αποτελούν ένα εργαλείο για την επεξεργασία δεδομένων που χρησιμοποιείται σε πολλές εφαρμογές. Είναι ένα σύνολο μεθόδων εκμάθησης με επίβλεψη, που χρησιμοποιούνται κυρίως στην ταξινόμηση και την παλινδρόμηση. Στην περίπτωση που δεν μπορεί να κατασκευαστεί ένα γραμμικό όριο, χρησιμοποιούνται οι συναρτήσεις πυρήνα, που μετατρέπουν το πρόβλημα σε γραμμικό. Χρησιμοποιούνται για την αναγνώριση μη φυσικών μοτίβων σε πολυμεταβλητές διεργασίες. Λόγω της αυξημένης ζήτησης της αποτελεσματικότητας της παραγωγής και της ποιότητας των προϊόντων, είναι απαραίτητα τα πολύπλοκα συστήματα ελέγχου, ώστε να ανιχνεύονται τα σφάλματα αυτόματα.

Οι περισσότερες διεργασίες παρακολούθησης και ελέγχου στην πραγματικότητα περιέχουν περισσότερες από μία μεταβλητές και συνήθως χρειάζεται ταυτόχρονη παρακολούθηση ή έλεγχος δύο ή περισσότερων χαρακτηριστικών. Σκοπός της παρούσας εργασίας είναι η παρουσίαση των SVM μεθόδων που χρησιμοποιούνται σε πολυμεταβλητές διεργασίες για τη βελτίωση τους, ώστε να αναγνωρίζονται μη φυσικά μοτίβα και να ανιχνεύονται τα σφάλματα εγκαίρως στα διαγράμματα ελέγχου. Επίσης γίνεται σύγκριση διαφόρων μεθοδολογιών για την παρακολούθηση διεργασιών με μη κανονικά δεδομένα, όπως το rk διάγραμμα, το AK διάγραμμα, το διάγραμμα ελέγχου μίας κλάσης και το K^2 διάγραμμα.

Συγκεκριμένα στο Κεφάλαιο 1 γίνεται μία παρουσίαση του Στατιστικού Ελέγχου Διεργασιών (SPC) για μονομεταβλητές και πολυμεταβλητές διεργασίες και το Κεφάλαιο 2 ασχολείται με τις Μηχανές Διανυσμάτων Υποστήριξης (SVM). Στο Κεφάλαιο 3 παρουσιάζεται η χρήση των SVM για την αναγνώριση μη φυσικών μοτίβων στα πολυμεταβλητά διαγράμματα ελέγχου ($MSPC$) και στο Κεφάλαιο 4 αναλύεται η αυτόματη ανίχνευση λαθών με τη χρήση των $SVMs$. Στο Κεφάλαιο 5 παρουσιάζονται μερικά διαγράμματα ελέγχου που βασίζονται στα $SVMs$ και στις τεχνικές του $MSPC$ και και τέλος στο Κεφάλαιο 6 παρουσιάζονται κάποιες εφαρμογές των μεθόδων που αναφέρθηκαν και τα αποτελέσματά τους.

Abstract

Statistical process control is one of the greatest technological developments of the twentieth century because it helps to achieve precision in process stability and improve capability through the reduction of variability. Statistical Process Control (SPC) detects on time the appearance of variability causes, in order to have corrective actions, before the construction of products, which do not meet the requirements. The most useful tool is the control chart, which improves the productivity, provides the flaws, prevents the unnecessary fitting and provides information about the capacity of the process.

Support vector machine (SVM) is a tool for data processing, which is used in many applications. It is a supervised statistical learning algorithm, which is used mostly in the classification and the regression. When we cannot construct a linear limit, we use the kernel functions, which convert the problem into a linear one. We use also SVMs for the recognition of non natural patterns in multivariate processes. Due to the increased request of effectiveness in production and the products' quality, it is necessary to use complex control systems, to detect the faults automatically.

Most of the monitoring and control processes contain several variables and usually we need simultaneous monitoring or control of two or more features. The purpose of this thesis is the presentation of the SVM methods, which are used in multivariate processes for improving them, in order to recognize non natural patterns and to detect the faults on time in the control charts. Moreover, we compare various methods for processes monitoring with non natural data, such as rk-chart, AK-chart, one class classification based chart and K^2 chart.

Specifically, Chapter 1 presents the Statistical Process Control for univariate and multivariate processes and in Chapter 2 deals with Support Vector Machines. In Chapter 3 is presented the use of SVMs for the recognition non-natural patterns in the multivariate control charts and in Chapter 4 we refer to the automatic fault detection with the use of SVMs. In the fifth Chapter are presented some control charts, which are based in the SVMs and in the MSPC techniques. Finally in Chapter 6 are presented some applications of the methods we refer to and their results.

ΚΑΤΑΛΟΓΟΣ ΣΥΝΤΟΜΟΓΡΑΦΙΩΝ

$ACUSUM$	Μη Παραμετρικό Πολυμεταβλητό Αθροιστικό Διάγραμμα Ελέγχου	Antirank-based Cumulative Sum Control Chart
AIR	Ακριβής Τιμή Αναγνώρισης	Accurate Identification Rates
AK	Προσαρμοσμένος Πυρήνας	Adaptive Kernel
ANN	Τεχνητό Νευρωνικό Δίκτυο	Artificial Neural Network
$ANOVA$	Ανάλυση της Διακύμανσης	Analysis Of Variance
ARL	Μέσο Μήκος Ροής	Average Run Length
ARL_0	Εντός Ελέγχου Μέσο Μήκος Ροής	In-Control Average Run Length
ARL_1	Εκτός Ελέγχου Μέσο Μήκος Ροής	Out-of-Control Average Run Length
ATS	Μέσος Χρόνος Σήματος	Average Time to Signal
BPN	Δίκτυο προς τα Πίσω Διάδοσης	Back Propagation Network
C_i^+	Άνω Συσσωρευμένο Άθροισμα	One Sided Upper Cusum
C_i^-	Κάτω Συσσωρευμένο Άθροισμα	One Sided Lower Cusum
CL	Κεντρική Γραμμή	Center Line
C_p	Δείκτης Ικανότητας Διεργασίας	Process Capability Ratio Index
$CUSUM$	Αθροιστικό Διάγραμμα Ελέγχου	Cumulative Sum Control Chart
DAG	Κατευθυνόμενο Ακυκλικό Γράφημα	Directed Acyclic Graph
$DICA$	Δυναμική Ανάλυση Ανεξάρτητων Συνιστωσών	Dynamic Independent Components Analysis
$EWMA$	Διάγραμμα Ελέγχου με Κινητούς Μέσους και Εκθετικά Βάρη	Exponentially Weighted Moving Average Control Chart
GS	Αναζήτηση Πλέγματος	Grid Search
ICA	Ανάλυση Ανεξάρτητων Συνιστωσών	Independent Components Analysis
IWL	Εσωτερικά Προειδοποιητικά Όρια	Inner Warning Limits
KDE	Εκτίμηση Πυκνότητας Πυρήνα	Kernel Density Estimation
$kNNDD$	Περιγραφή Δεδομένων των κ Κοντινότερων Γειτόνων	k Nearest Neighbours Data Description
LCL	Κάτω Όρια Ελέγχου	Lower Control Limits
LSL	Κάτω Όρια Προδιαγραφών	Lower Specification Limits
$LS-SVM$	Μηχανές Διανυσματικής Υποστήριξης Ελάχιστων Τετραγώνων	Least Squares Support Vector Machines

LVQ	Δίκτυα Εκμάθησης Κβαντικού Διανύσματος	Learning Vector Quantization
$MARS$	Προσαρμοσμένα πολυμεταβλητά splines παλινδρόμησης	Multivariate Adaptive Regression Splines
$MCUSUM$	Πολυμεταβλητό Αθροιστικό Διάγραμμα Ελέγχου	Multivariate Cumulative Sum Control Chart
$MEWMA$	Πολυμεταβλητό Διάγραμμα Ελέγχου με Κινητούς Μέσους και Εκθετικά Βάρη	Multivariate Exponentially Weighted Moving Average Control Chart
MLP	Πολυστρωματικοί Αισθητήρες	Multi−Layer Perceptron
MR	Κινούμενο Εύρος	Moving Range
$MSPC$	Πολυμεταβλητός Στατιστικός Έλεγχος Διεργασιών	Multivariate Statistical Process Control
$MSPM$	Παρακολούθηση Πολυμεταβλητής Στατιστικής Διεργασίας	Multivariate Statistical Process Monitoring
OAA	Ένας Εναντίον Όλων μέθοδος	One against All method
OAO	Ένας Εναντίον Ενός μέθοδος	One against One method
$OCAP$	Εκτός Ελέγχου Πρόγραμμα Δράσης	Out of Control Action Plan
$OC-SVM$	Διάγραμμα Ελέγχου Μίας Κλάσης	One-class Classification Control Chart
OWL	Εξωτερικά Προειδοποιητικά Όρια	Outer Warning Limits
PCA	Ανάλυση Κύριων Συνιστωσών	Principal Component Analysis
PSO	Βελτιστοποίηση Πλήθους Σωματιδίων	Particle Swarm Optimization
RBF	Συναρτήσεις Ακτινικής Βάσης	Radial Basis Function
$ROCC$	Βαθμός της Σωστής Ταξινόμησης	Rate Of Correct Classification
SPC (ΣΕΔ)	Στατιστικός Έλεγχος Διεργασιών	Statistical Process Control
SPE	Τετραγωνικό Σφάλμα Πρόβλεψης	Squared Prediction Error
SR	Σήμα Αντίδρασης	Signal Resistance
$SVDD$	Περιγραφή Δεδομένων Διανυσματικής Υποστήριξης	Support Vector Data Description
SVM	Μηχανή Διανυσμάτων Υποστήριξης	Support Vector Machine
$SVRM$	Μηχανή Αναπαράστασης Διανυσμάτων Υποστήριξης	Support Vector Representation Machine
UCL	Άνω Όρια Ελέγχου	Upper Control Limits
USL	Άνω Όρια Προδιαγραφών	Upper Specification Limits
WA	Κυματική Ανάλυση	Wavelet Analysis

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

1 Εισαγωγή στον Στατιστικό Έλεγχο Διεργασιών

1.1 Στατιστικός Έλεγχος Διεργασιών

Ο Στατιστικός Έλεγχος Διεργασιών (ΣΕΔ) περιλαμβάνει εργαλεία επίλυσης προβλημάτων χρήσιμα στην επίτευξη ακρίβειας σταθερότητας της διεργασίας και στη βελτίωση της ικανότητας της διεργασίας μειώνοντας τη μεταβλητότητα. Είναι μία από τις μεγαλύτερες τεχνολογικές ανακαλύψεις του 20ου αιώνα γιατί

- βασίζεται σε θεμελιώδεις αρχές,
- είναι εύκολος στη χρήση,
- έχει σημαντική επίδραση και
- εφαρμόζεται σε κάθε διεργασία.

Περιλαμβάνει 7 κύρια εργαλεία τα οποία ονομάζονται "$magnificent\ seven$" και είναι τα εξής:

- Ιστόγραμμα, φυλλόγραμμα *(Stem and leaf plot)*
- Φύλλο ελέγχου *(Check sheet)*
- Διάγραμμα *Pareto (Pareto chart)*
- Διάγραμμα αιτίου και αποτελέσματος *(Cause and effect diagram)*
- Διάγραμμα συγκέντρωσης ατελειών *(Defect concentration diagram)*
- Διάγραμμα διασποράς *(Scatter diagram)*
- Διάγραμμα ελέγχου *(Control chart).*

Τα 7 αυτά εργαλεία αποτελούν ένα πολύ σημαντικό μέρος του ΣΕΔ, με βασικότερο να είναι το διάγραμμα ελέγχου, το οποίο βοηθάει στην παρακολούθηση μιας παραγωγικής διεργασίας. Η ορθή ανάπτυξη του ΣΕΔ βοηθάει στη δημιουργία ενός περιβάλλοντος, στο οποίο όλοι ατομικά σε μία οργάνωση επιδιώκουν συνεχή βελτίωση στην ποιότητα και στην παραγωγή.

Σε κάθε παραγωγική διεργασία υπάρχει πάντα μία φυσική μεταβλητότητα, ανεξάρτητα από το πόσο καλά σχεδιασμένη είναι, η οποία οφείλεται σε πολλά, μικρά, αναπόφευκτα αίτια. Αυτά καλούνται **τυχαία αίτια μεταβλητότητας** (*chance causes of variation*) και η διεργασία η οποία λειτουργεί με την

παρουσία φυσικής μεταβλητότητας καλείται **εντός στατιστικού ελέγχου** (*in statistical control*).

Άλλα είδη μεταβλητότητας που μπορεί να εμφανιστούν σε μία διεργασία οφείλονται σε λανθασμένα ρυθμισμένες μηχανές, σε λάθος του χειριστή κάποιου μηχανήματος ή σε ελαττωματική πρώτη ύλη. Η μεταβλητότητα που οφείλεται σε αυτά τα αίτια είναι σημαντικά μεγαλύτερη από τη φυσική και συνήθως οδηγεί σε μη αποδεκτά επίπεδα λειτουργίας της παραγωγικής διεργασίας. Αυτή η μεταβλητότητα καλείται ειδική και τα αίτια που οδηγούν σε αυτή **ειδικά ή προσδιορισμένα αίτια** (*special or assignable causes*) **μεταβλητότητας**. Η διεργασία η οποία λειτουργεί με την παρουσία ειδικής μεταβλητότητας καλείται **εκτός στατιστικού ελέγχου** (*out of statistical control*) ή ότι λειτουργεί σε **ασταθή κατάσταση** (*unstable state*).

Συνήθως οι διεργασίες βρίσκονται σε στατιστικό έλεγχο για σχετικά μεγάλο χρονικό διάστημα. Ωστόσο, καμία διεργασία δεν μπορεί να είναι σταθερή για πάντα, καθώς κάποια στιγμή θα εμφανιστούν ειδικά αίτια που θα οδηγήσουν σε μία αλλάγη της κατάστασης της διεργασίας σε εκτός ελέγχου. Επίσης, υπάρχουν τα άνω και κάτω όρια προδιαγραφών (*upper and lower specification limits*, USL *and* LSL), αντίστοιχα, τα οποία καθορίζονται στη φάση σχεδιασμού και μας ενημερώνουν πότε η διεργασία βρίσκεται εκτός ελέγχου. Όταν οι τιμές του ποιοτικού χαρακτηριστικού σχεδιαστούν εκτός αυτών των ορίων τότε γνωρίζουμε ότι η διεργασία είναι εκτός ελέγχου και πρέπει να ερευνηθούν τα αίτια. Κύριο αντικείμενο του ΣΕΔ είναι να ανιχνεύσει έγκαιρα την εμφάνιση ειδικών αιτιών μεταβλητότητας σε μία διεργασία έτσι ώστε να ερευνήσουμε τα αίτια και να προβούμε στις απαραίτητες διορθωτικές ενέργειες προτού κατασκευαστούν προϊόντα που δεν πληρούν τις προδιαγραφές. Το διάγραμμα ελέγχου είναι μία μέθοδος η οποία σε πραγματικό χρόνο ανιχνεύει τα διάφορα αίτια μεταβλητότητας σε μία διεργασία.

1.2 Διαγράμματα ελέγχου

1.2.1 Εισαγωγή και βασικές αρχές

Το διάγραμμα ελέγχου εφευρέθηκε από τον Walter Shewhart και χρησιμοποιείται ευρέως για την ανίχνευση αιτιών που προκαλούν μεταβλητότητα σε μία διεργασία, για την εκτίμηση των παραμέτρων της διεργασίας παραγωγής και για τον καθορισμό της ικανότητας της διεργασίας. Επίσης παρέχει χρήσιμες πληροφορίες για την βελτίωση της διεργασίας.

Είναι μία γραφική αναπαράσταση ενός ποιοτικού χαρακτηριστικού το οποίο έχει μετρηθεί πειραματικά ή έχει υπολογιστεί από ένα δείγμα σε συνάρτηση με το χρόνο. Περιέχει μία κεντρική γραμμή (*center line*, CL) η οποία αναπαριστά

τη μέση τιμή του ποιοτικού χαρακτηριστικού όταν η διεργασία βρίσκεται εντός στατιστικού ελέγχου και τα άνω και κάτω όρια ελέγχου (*Upper and Lower Control Limits*, UCL *and* LCL), αντίστοιχα, τα οποία έχουν επιλεχθεί έτσι ώστε όταν η διεργασία βρίσκεται εντός στατιστικού ελέγχου, όλα τα σημεία του δείγματος να βρίσκονται εντός αυτών. Μία άλλη ένδειξη, ότι η διεργασία βρίσκεται εκτός ελέγχου, είναι τα σημεία που δεν είναι τυχαία κατανεμημένα. Για να είναι η διεργασία εντός ελέγχου πρέπει όλα τα σημεία να είναι κατανεμημένα εντός των ορίων με τυχαίο τρόπο. Συνήθως αυτά τα σημεία ενώνονται με μία τεθλασμένη γραμμή έτσι ώστε να απεικονίζεται η εξέλιξη της διεργασίας στο χρόνο. Κατά το σχεδιασμό του διαγράμματος ελέγχου πρέπει να γίνει επίσης η επιλογή του μεγέθους του δείγματος και να οριστεί η συχνότητα δειγματοληψίας.

Τα διαγράμματα ελέγχου είναι δημοφιλή κυριώς γιατί είναι μία αποδεδειγμένη τεχνική για τη βελτίωση της παραγωγικότητας, είναι αποτελεσματικά στην πρόβλεψη των ελαττωμάτων, αποτρέπουν την άσκοπη διαδικασία προσαρμογής, παρέχουν διαγνωστικές πληροφορίες και πληροφορίες σχετικές με την ικανότητα της διεργασίας. Ένα διάγραμμα ελέγχου πρέπει να συνοδεύεται από ένα εκτός ελέγχου πρόγραμμα δράσης (*out of control action plan*, $OCAP$), το οποίο θα ενεργοποιείται κάθε φορά που το διάγραμμα ελέγχου δείχνει ειδικά αίτια μεταβλητότητας στη διαδικασία.

Τα διαγράμματα ελέγχου διακρίνονται σε δύο κατηγορίες ανάλογα με το είδος της μεταβλητής που περιγράφει ένα ποιοτικό χαρακτηριστικό του προϊόντος:

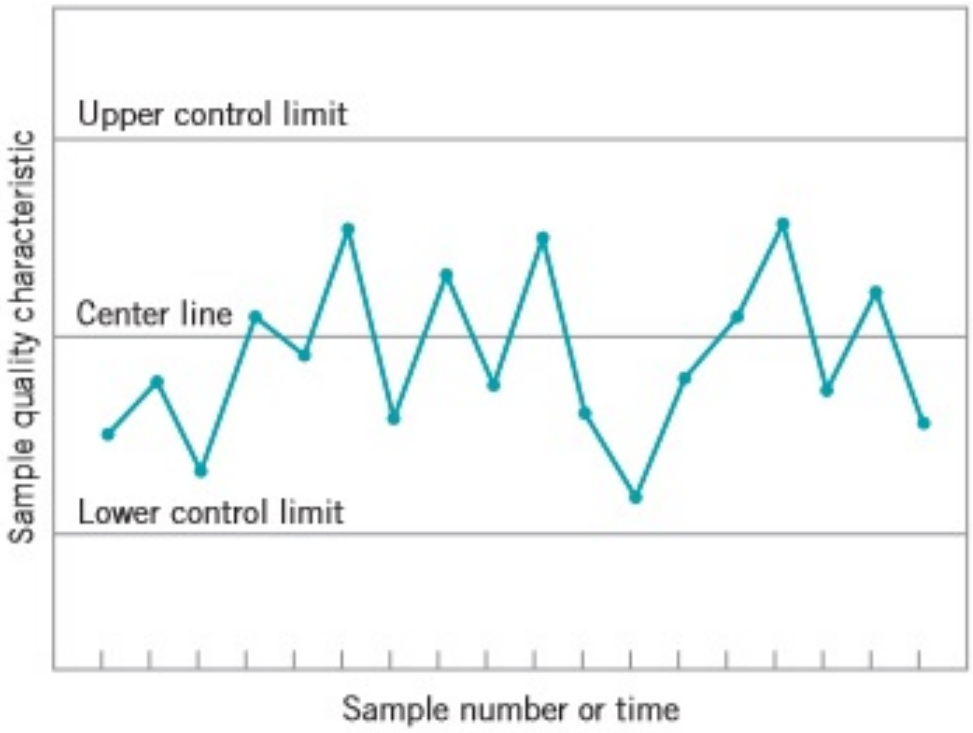
- Διαγράμματα ελέγχου για συνεχείς μεταβλητές (*control charts for variables*)
- Διαγράμματα ελέγχου για διακριτά χαρακτηριστικά (*control charts for attributes*)

Παρακάτω δίνεται ένα γενικό μοντέλο για τα διαγράμματα ελέγχου:

$$\begin{aligned} UCL &= \mu_W + L\,\sigma_W \\ CL &= \mu_W \\ LCL &= \mu_W - L\,\sigma_W \end{aligned}$$

όπου μ_W είναι η μέση τιμή και σ_W η τυπική απόκλιση της στατιστικής συνάρτησης W που απεικονίζεται στο διάγραμμα ελέγχου, η οποία μετρά κάποιο ποιοτικό χαρακτηριστικό που μας ενδιαφέρει και L είναι η απόσταση των ορίων ελέγχου από την κεντρική γραμμή.

Το διάγραμμα ελέγχου τύπου *Shewhart*, το οποίο χρησιμοποιείται ευρέως λόγω της απλότητας στην κατασκευή και στην ερμηνεία του, είναι ένα μοντέλο



Σχήμα 1.1: Ένα τυπικό διάγραμμα ελέγχου

ορίων σ. Τα διαγράμματα ελέγχου *Shewhart* είναι πιο αποτελεσματικά όταν τα δεδομένα είναι ασυσχέτιστα με στατική συμπεριφορά, οπότε και η αναπαράστασή τους είναι τυχαία, και οι προηγούμενες τιμές δεν δίνουν στοιχεία για να προβλέψουμε τις μελλοντικές. Τα διαγράμματα ελέγχου μπορούν να κατασκευαστούν έτσι ώστε η απόδοση της διεργασίας να είναι προβλέψιμη και λογική στο χρήστη και να είναι αποτελεσματικά στο να διακρίνουν καταστάσεις εκτός ελέγχου.

1.2.2 Επιλογή ορίων ελέγχου

Η επιλογή των ορίων ελέγχου σε ένα διάγραμμα ελέγχου είναι μία από τις πιο κρίσιμες αποφάσεις. Αν είναι πολύ μακριά από την κεντρική γραμμή μειώνεται η πιθανότητα ένα σημείο να σχεδιαστεί εκτός των ορίων ενώ η κατάσταση είναι εκτός ελέγχου και αν σχεδιαστούν πολύ κοντά στην κεντρική γραμμή αυξάνεται αυτή η πιθανότητα.

Υπάρχουν επίσης και τα προειδοποιητικά όρια ελέγχου τα οποία χρησιμο-

ποιούνται για να ανιχνεύονται πιο έγκαιρα τα ειδικά αίτια μεταβλητότητας σε μία διεργασία. Ωστόσο έχουν το μειονέκτημα ότι πολλές φορές μπερδέυουν το χρήστη και αυξάνουν το ρίσκο για λανθασμένη προειδοποίηση. Διακρίνονται σε εσωτερικά προειδοποιητικά όρια (*inner warning limits*) και σε εξωτερικά (*outer warning limits*). Τα εσωτερικά σχεδιάζονται σε απόσταση σ από την κεντρική γραμμή του διαγράμματος και τα εξωτερικά σε απόσταση 2σ από την κεντρική γραμμή του διαγράμματος. Αν ένα ή περισσότερα σημεία βρίσκονται ανάμεσα στα εξωτερικά προειδοποιητικά όρια και στα όρια ελέγχου, έχουμε ενδείξεις ότι η διεργασία βρίσκεται εκτός ελέγχου.

1.2.3 Μέτρα απόδοσης ενός διαγράμματος ελέγχου

Το μέσο μήκος ροής (*average run length - ARL*) του διαγράμματος είναι ένα μέτρο απόδοσής του. Είναι ο μέσος αριθμός των σημείων που πρέπει να σχεδιαστούν στο διάγραμμα ελέγχου προτού εμφανιστεί ένα σημείο εκτός των ορίων ελέγχου. Αν οι παρατηρήσεις είναι ασυσχέτιστες, τότε το ARL για κάθε διάγραμμα ελέγχου *Shewhart* μπορεί να υπολογιστεί από τον τύπο:

$$ARL = \frac{1}{p}, \tag{1.1}$$

όπου p είναι η πιθανότητα ένα σημείο να σχεδιαστεί εκτός των ορίων ελέγχου.

Αν α είναι η πιθανότητα ένα σημείο να βρεθεί εκτός των ορίων ενώ η διεργασία βρίσκεται εντός ελέγχου και το χαρακτηριστικό του προϊόντος $X \sim N(\mu, \sigma^2)$, τότε $ARL_0 = \frac{1}{a}$. Το μέτρο αυτό καλείται εντός ελέγχου μέσο μήκος ροής ($in-control\ average\ run\ length$).

Αντίθετα όταν η διεργασία είναι εκτός ελέγχου, το μέσο μήκος ροής υπολογίζεται από τον τύπο: $ARL_1 = \frac{1}{1-\beta}$, όπου β είναι η πιθανότητα ένα σημείο να βρεθεί εντός των ορίων ελέγχου. Το μέτρο αυτό καλείται εκτός ελέγχου μέσο μήκος ροής ($out-of-control\ average\ run\ length$).

Επίσης μπορούμε να βρούμε τον μέσο χρόνο σήματος (*average time to signal*, ATS), έχοντας τα καθορισμένα χρονικά διαστήματα των ωρών στα οποία παίρνουμε δείγμα, έστω h, τότε $ATS = ARL * h$.

1.2.4 Κανόνες ευαισθητοποίησης για τα διαγράμματα ελέγχου

Το βασικό κριτήριο για να θεωρήσουμε μία κατάσταση εκτός ελέγχου είναι να εμφανιστούν σημεία εκτός των ορίων ελέγχου. Υπάρχουν όμως και κάποια συ-

μπληρωματικά κριτήρια που αυξάνουν την ευαισθησία των διαγραμμάτων ελέγχου, όπως να εμφανιστεί στο διάγραμμα κάποιο συγκεκριμένο μοτίβο *(pattern)*. Οι σημαντικότεροι κανόνες που χρησιμοποιούνται για την ευαισθητοποίηση ενός διαγράμματος ελέγχου είναι οι εξής:

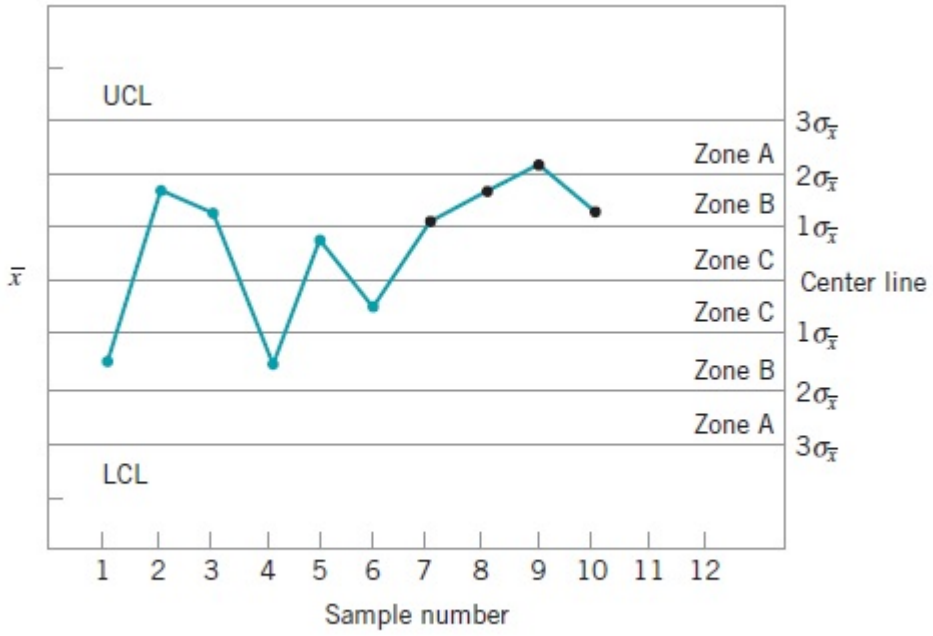
1. Ένα ή περισσότερα σημεία να βρίσκονται εκτός των 3σ ορίων ελέγχου.
2. Δύο στα τρία συνεχόμενα σημεία να βρίσκονται στη ζώνη Α.
3. Τέσσερα στα πέντε συνεχόμενα σημεία να βρίσκονται εκτός της ζώνης C.
4. Οκτώ συνεχόμενα σημεία να βρίσκονται από τη μία πλευρά της κεντρικής γραμμής.
5. Έξι συνεχόμενα σημεία να βρίσκονται σε αύξουσα ή φθίνουσα διάταξη.
6. Δεκαπέντε σημεία στη σειρά να βρίσκονται στη ζώνη C.
7. Δεκατέσσερα σημεία στη σειρά να εναλλάσσονται της κεντρικής γραμμής.
8. Οκτώ συνεχόμενα σημεία να βρίσκονται εκτός της ζώνης C.
9. Να υπάρχει ένα μη τυχαίο ή ασυνήθιστο μοτίβο στα δεδομένα.
10. Ένα ή περισσότερα σημεία να βρίσκονται κοντά στα όρια ή στα προειδοποιητικά όρια ελέγχου.

Οι κανόνες 1 - 4 είναι γνωστοί ως *Western Electric Rules* (Σχήμα 1.2).

1.2.5 Αναγνώριση μοτίβων

Στην προήγουμενη ενότητα αναφερθήκαμε σε μη τυχαία μοτίβα. Υπάρχουν διάφοροι τύποι μοτίβων οι οποίοι δεν είναι τυχαίοι, όπως

- Τα κυκλικά *(cyclic)* μοτίβα, όπου υπάρχουν ημιτονοειδής μορφές με επανειλημμένες ανοδικές και καθοδικές τάσεις στα δεδομένα.
- Τα συστηματικά (*systematic*) μοτίβα, όπου η σειρά των σημείων στο διάγραμμα είναι προβλέψιμη με ένα συστηματικό τρόπο.



Σχήμα 1.2: Οι κανόνες *Western Electric*, όπου τα 4 τελευταία σημεία παραβιάζουν τον κανόνα 3.

- Τα μοτίβα ανοδικής και καθοδικής μεταβολής *(upward και downward shift, αντίστοιχα)*, όπου υπάρχουν απότομες αλλαγές από χαμηλό σε υψηλό επίπεδο και από υψηλό σε χαμηλό επίπεδο, αντίστοιχα.

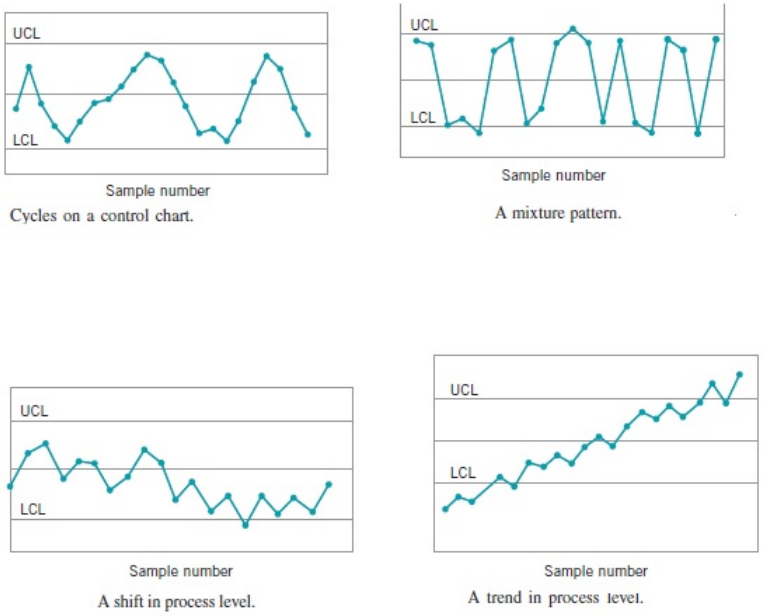
- Τα μοτίβα αυξανόμενης και μειούμενης τάσης *(increasing και decreasing trend, αντίστοιχα)*, όπου υπάρχει μία σειρά σημείων που έχουν μια βαθμιαία ανοδική και καθοδική μεταβολή αντίστοιχα στο μέσο όρο.

- Η στρωμάτωση *(stratification)* ή η τάση των σημείων να συλλέγονται τεχνικά γύρω από την κεντρική γραμμή. Ύπάρχει μία έλλειψη φυσικής μεταβλητότητας στα παρατηρούμενα μοτίβα.

  Τα προαναφερθέντα μοτίβα παρουσιάζονται στο Σχήμα 1.3.

1.2.6 Φάσεις I και II

Στη φάση I συλλέγονται τα δεδομένα και αναλύονται για να καθοριστεί αν η διεργασία είναι εντός ή εκτός ελέγχου κατά τη διάρκεια συλλογής των δε-



Σχήμα 1.3: Τυχαία μοτίβα

δομένων. Κατασκευάζονται δοκιμαστικά όρια ελέγχου, τα οποία βοηθούν το χρήστη να φέρει τη διεργασία εντός ελέγχου. Όταν η διεργασία γίνει εντός ελέγχου, το διάγραμμα ελέγχου που προκύπτει είναι κατάλληλο για την παρακολούθηση της μελλοντικής συμπεριφοράς της διεργασίας.

Κατά τη διάρκεια της φάσης II, οποία λέγεται και φάση παρακολούθησης, τα δεδομένα βρίσκονται κάτω από μία σταθερή κατάσταση και το διάγραμμα ελέγχου χρησιμοποιείται έτσι ώστε να ελέγχεται αν η διαδικασία είναι εντός ελέγχου. Έτσι μπορεί ο χρήστης να ανιχνεύσει έγκαιρα μια αλλαγή στο μέσο επίπεδο των χαρακτηριστικών που καθορίζουν την ποιότητα του παραγόμενου προϊόντος. Σε αυτή τη φάση είναι χρήσιμο το ARL, για να αξιολογήσουμε την απόδοση του διαγράμματος ελέγχου.

1.3 Τα υπόλοιπα από τα 7 κύρια εργαλεία του Στατιστικού Ελέγχου Διεργασιών

Τα διαγράμματα ελέγχου είναι μεν ένα χρήσιμο εργαλείο που λύνει προβλήματα και βελτιώνει τη διεργασία αλλά είναι πιο αποτελεσματικό, όταν η χρήση του είναι πλήρως ολοκληρωμένη σε ένα ΣΕΔ. Τα επτά αυτά εργαλεία επίλυσης προβλημάτων χρησιμοποιούνται για την αναγνώριση ευκαιριών βελτίωσης,

στη βοήθεια μείωσης της μεταβλητότητας και στον εκμηδενισμό των άχρηστων προϊόντων.

- Φύλλο ελέγχου: Γίνεται στην αρχή της βελτίωσης της διεργασίας και περιλαμβάνει τη συλλογή του ιστορικού ή των πρόσφατων δεδομένων λειτουργίας της διεργασίας. Κατά το σχεδιασμό του είναι σημαντικό να καθοριστεί ο τύπος των δεδομένων που συλλέγονται, ο αριθμός της διεργασίας, η ημερομηνία, τα στοιχεία του αναλυτή και άλλες πληροφορίες χρήσιμες για τη διάγνωση αιτιών κακής απόδοσης.

- Διάγραμμα *Pareto*: Είναι η κατανομή συχνοτήτων των ιδιοτήτων των δεδομένων ανά κατηγορία. Μέσω αυτού ο χρήστης μπορεί γρήγορα να αναγνωρίζει το πιο συχνά εμφανιζόμενο τύπο σφαλμάτων, όχι όμως και το πιο σημαντικό. Υπάρχουν όμως παραλλαγές αυτού που να δείχνουν το πιο σημαντικό αίτιο. Χρησιμοποιείται ευρέως σε μη-κατασκευαστικές εφαρμογές σε διαδικασίες βελτίωσης ποιότητας.

- Διάγραμμα αιτίου και αποτελέσματος: Είναι ένα εργαλείο που χρησιμοποιείται συχνά για να αναγνωριστούν και να απομονωθούν πιθανά αίτια σφαλμάτων, κυρίως μέσω της συζήτησης.

- Διάγραμμα συγκέντρωσης ατελειών: Καθορίζει αν η θέση των ελαττωμάτων στο διάγραμμα περιέχει κάποια χρήσιμη πληροφορία για τα πιθανά αίτια.

- Διάγραμμα διασποράς: Αναγνωρίζει τη συσχέτιση μεταξύ δύο μεταβλητών και είναι πολύ χρήσιμο στη μοντελοποίηση της ανάλυσης.

1.4 Διαγράμματα ελέγχου μεταβλητών

1.4.1 Εισαγωγή

Ένα ποιοτικό χαρακτηριστικό που μετράται σε αριθμητική κλίμακα, το καλούμε μεταβλητή. Απλά μετρήσιμα χαρακτηριστικά είναι το μήκος, το πλάτος, η θερμοκρασία ή η ένταση. Τα διαγράμματα ελέγχου μεταβλητών χρησιμοποιούνται τόσο στη βιομηχανία όσο και σε πολλές επιστημονικές περιοχές. Ο έλεγχος του δειγματικού μέσου γίνεται με το διάγραμμα ελέγχου για το μέσο, όπως το $\bar{x}$ διάγραμμα. Ο έλεγχος της διασποράς της διεργασίας μπορεί να ελεγχθεί είτε με το διάγραμμα ελέγχου για την τυπική απόκλιση (S) ή με ένα διάγραμμα ελέγχου για το εύρος (R).

1.4.2 $\bar{x}$ και R διαγράμματα ελέγχου

Το $\bar{x}$ διάγραμμα ελέγχου χρησιμοποιείται για κάθε ποιοτικό χαρακτηριστικό που μας ενδιαφέρει. Έστω x το χαρακτηριστικό του προϊόντος που μας ενδιαφέρει, τέτοιο ώστε $x \sim N(\mu, \sigma^2)$ με μ και σ γνωστά. Αν τα x_1,...,x_n είναι ένα δείγμα μεγέθους n, τότε ο δειγματικός μέσος δίνεται από τον τύπο

$$\bar{x} = \frac{x_1 + ... + x_n}{n} \tag{1.2}$$

και ακολουθεί κανονική κατανομή με μέση τιμή μ και τυπική απόκλιση $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ και η πιθανότητα οποιαδήποτε τιμή του μέσου να βρίσκεται στο διάστημα $[\mu - z_{\alpha/2}\sigma_x \ , \ \mu + z_{\alpha/2}\sigma_x]$ είναι 1-α.
Τα όρια ελέγχου του διαγράμματος για το δειγματικό μέσο είναι:

$$\begin{aligned} UCL &= \mu + 3\sigma_x \\ CL &= \mu \\ LCL &= \mu - 3\sigma_x. \end{aligned}$$

Όταν οι τιμές των μ και σ είναι άγνωστες πρέπει να εκτιμηθούν και για να γίνει αυτό επιλέγονται $m = 20$ ως 25 ανεξάρτητα τυχαία δείγματα μεγέθους $n = 4$ ως 6 το καθένα, υποθέτοντας ότι η επιλογή των δειγμάτων έγινε όταν η διεργασία ήταν εντός ελέγχου.
Το μ εκτιμάται από τον τύπο

$$\hat{\mu} = \bar{\bar{x}} = \frac{1}{m}\sum_{i=1}^{m} \bar{x}_i. \tag{1.3}$$

Το σ εκτιμάται με 2 τρόπους

1. **Από το R:**
   Έστω $R_1, \ldots, R_m$ τα εύρη των m δειγμάτων με $\mu_{R_i} = E(R_i) = \sigma d_2$ και $\sigma_{R_i} = \sqrt{Var(R_i)} = \sigma d_3$. Θέτοντας $\bar{R} = \frac{1}{m}(R_1 + R_2 + \ldots + R_m)$ προκύπτει $E(\bar{R}) = \sigma d_2$. Δηλαδή $\hat{\sigma} = \bar{R}/d_2$. Τα d_2 και d_3 είναι σταθερές και εξαρτώνται από το μέγεθος του δείγματος (n).

2. **Από το S:**
   Έστω:
   $$S_i = \sqrt{S_i^2} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2} \tag{1.4}$$
   με $\mu_{S_i} = E(S_i) = \sigma c_4$ και $\sigma_{S_i} = \sqrt{Var(S_i)} = \sigma\sqrt{1 - c_4^2}$, όπου η c_4 σταθερά εξαρτάται από το μέγεθος n του δείγματος.

Αν θέσουμε $\bar{S}=\frac{1}{m}(S_1+S_2+\ldots S_m)$ τότε έχουμε $E(\bar{S})=\sigma c_4$.
Δηλαδή $\hat{\sigma}=\bar{S}/c_4$.

Τα όρια ελέγχου για το $\bar{x}$ διάγραμμα είναι:

$$\begin{aligned} UCL &= \bar{\bar{x}} + A_2\bar{R} \\ CL &= \bar{\bar{x}} \\ LCL &= \bar{\bar{x}} - A_2\bar{R}. \end{aligned}$$

Και για το R διάγραμμα ελέγχου είναι:

$$\begin{aligned} UCL &= D_4\bar{R} \\ CL &= \bar{R} \\ LCL &= D_3\bar{R}. \end{aligned}$$

Οι σταθερές $A_2 = \frac{3}{d_2\sqrt{n}}$, $D_3 = 1 - 3\frac{d_3}{d_2}$, $D_4 = 1 + 3\frac{d_3}{d_2}$, d_2, d_3, c_4 δίνονται στον Πίνακα 1.1.

Στη φάση Ι της κατασκευής του διαγράμματος ελέγχου, όταν χρησιμοποιούνται τα πρώτα δείγματα για την κατασκευή των $\bar{x}$ και R διαγραμμάτων ελέγχου, θεωρούμε τα δοκιμαστικά όρια. Για να καθορίσουμε αν η διαδικασία είναι εντός ελέγχου, σχεδιάζουμε τις τιμές $\bar{x}$ και R για κάθε δείγμα και αναλύουμε την απεικόνιση. Αν είναι όλα τα σημεία εντός των ορίων και δεν υπάρχει κάποιο συστηματικό μοτίβο, θεωρούμε τη διεργασία εντός ελέγχου και τα όρια ελέγχου κατάλληλα για μελλοντική χρήση. Αν υπάρχουν ενδείξεις ότι η διεργασία είναι εκτός ελέγχου, χρειάζεται επανεξέταση των ορίων ελέγχου, μέσω έρευνας για ειδικά αίτια. Αν βρεθούν ειδικά αίτια τα εξαλείφουμε και επαναϋπολογίζουμε τα όρια ελέγχου με τα εναπομείναντα σημεία. Αυτή η διαδικασία συνεχίζεται μέχρι η διεργασία να βρεθεί εντός ελέγχου.

1.4.2.1 Εκτίμηση της ικανότητας της μεθόδου

Από το $\bar{x}$ διάγραμμα ελέγχου μπορούμε να εκτιμήσουμε το $\bar{\bar{x}}$ και από το R διάγραμμα την τυπική απόκλιση $\hat{\sigma}$.
Ο δείκτης ικανότητας της διεργασίας (*process capability ratio/index*) για ένα ποιοτικό χαρακτηριστικό είναι: $C_p=\frac{USL-LSL}{6\sigma}$.

Πίνακας 1.1: Οι τιμές των σταθερών για διάφορες τιμές του n

Factors for Constructing Variables Control Charts

Observations in Sample, n	Chart for Averages: Factors for Control Limits A	A_2	A_3	Chart for Averages: Factors for Center Line c_4	$1/c_4$	Chart for Standard Deviations: Factors for Control Limits B_3	B_4	B_5	B_6	Factors for Center Line d_2	$1/d_2$	Chart for Ranges: Factors for Control Limits d_3	D_1	D_2	D_3	D_4
2	2.121	1.880	2.659	0.7979	1.2533	0	3.267	0	2.606	1.128	0.8865	0.853	0	3.686	0	3.267
3	1.732	1.023	1.954	0.8862	1.1284	0	2.568	0	2.276	1.693	0.5907	0.888	0	4.358	0	2.574
4	1.500	0.729	1.628	0.9213	1.0854	0	2.266	0	2.088	2.059	0.4857	0.880	0	4.698	0	2.282
5	1.342	0.577	1.427	0.9400	1.0638	0	2.089	0	1.964	2.326	0.4299	0.864	0	4.918	0	2.114
6	1.225	0.483	1.287	0.9515	1.0510	0.030	1.970	0.029	1.874	2.534	0.3946	0.848	0	5.078	0	2.004
7	1.134	0.419	1.182	0.9594	1.0423	0.118	1.882	0.113	1.806	2.704	0.3698	0.833	0.204	5.204	0.076	1.924
8	1.061	0.373	1.099	0.9650	1.0363	0.185	1.815	0.179	1.751	2.847	0.3512	0.820	0.388	5.306	0.136	1.864
9	1.000	0.337	1.032	0.9693	1.0317	0.239	1.761	0.232	1.707	2.970	0.3367	0.808	0.547	5.393	0.184	1.816
10	0.949	0.308	0.975	0.9727	1.0281	0.284	1.716	0.276	1.669	3.078	0.3249	0.797	0.687	5.469	0.223	1.777
11	0.905	0.285	0.927	0.9754	1.0252	0.321	1.679	0.313	1.637	3.173	0.3152	0.787	0.811	5.535	0.256	1.744
12	0.866	0.266	0.886	0.9776	1.0229	0.354	1.646	0.346	1.610	3.258	0.3069	0.778	0.922	5.594	0.283	1.717
13	0.832	0.249	0.850	0.9794	1.0210	0.382	1.618	0.374	1.585	3.336	0.2998	0.770	1.025	5.647	0.307	1.693
14	0.802	0.235	0.817	0.9810	1.0194	0.406	1.594	0.399	1.563	3.407	0.2935	0.763	1.118	5.696	0.328	1.672
15	0.775	0.223	0.789	0.9823	1.0180	0.428	1.572	0.421	1.544	3.472	0.2880	0.756	1.203	5.741	0.347	1.653
16	0.750	0.212	0.763	0.9835	1.0168	0.448	1.552	0.440	1.526	3.532	0.2831	0.750	1.282	5.782	0.363	1.637
17	0.728	0.203	0.739	0.9845	1.0157	0.466	1.534	0.458	1.511	3.588	0.2787	0.744	1.356	5.820	0.378	1.622
18	0.707	0.194	0.718	0.9854	1.0148	0.482	1.518	0.475	1.496	3.640	0.2747	0.739	1.424	5.856	0.391	1.608
19	0.688	0.187	0.698	0.9862	1.0140	0.497	1.503	0.490	1.483	3.689	0.2711	0.734	1.487	5.891	0.403	1.597
20	0.671	0.180	0.680	0.9869	1.0133	0.510	1.490	0.504	1.470	3.735	0.2677	0.729	1.549	5.921	0.415	1.585
21	0.655	0.173	0.663	0.9876	1.0126	0.523	1.477	0.516	1.459	3.778	0.2647	0.724	1.605	5.951	0.425	1.575
22	0.640	0.167	0.647	0.9882	1.0119	0.534	1.466	0.528	1.448	3.819	0.2618	0.720	1.659	5.979	0.434	1.566
23	0.626	0.162	0.633	0.9887	1.0114	0.545	1.455	0.539	1.438	3.858	0.2592	0.716	1.710	6.006	0.443	1.557
24	0.612	0.157	0.619	0.9892	1.0109	0.555	1.445	0.549	1.429	3.895	0.2567	0.712	1.759	6.031	0.451	1.548
25	0.600	0.153	0.606	0.9896	1.0105	0.565	1.435	0.559	1.420	3.931	0.2544	0.708	1.806	6.056	0.459	1.541

For $n > 25$.

$$A = \frac{3}{\sqrt{n}} \quad A_3 = \frac{3}{c_4\sqrt{n}} \quad c_4 \cong \frac{4(n-1)}{4n-3}$$

$$B_3 = 1 - \frac{3}{c_4\sqrt{2(n-1)}} \quad B_4 = 1 + \frac{3}{c_4\sqrt{2(n-1)}}$$

$$B_5 = c_4 - \frac{3}{\sqrt{2(n-1)}} \quad B_6 = c_4 + \frac{3}{\sqrt{2(n-1)}}$$

- Αν $C_p > 1$ παράγονται λίγα σημεία εκτός των ορίων ελέγχου.

- Αν $C_p = 1$ χρησιμοποιείται όλη η ζώνη ανοχής της διεργασίας.

- Αν $C_p < 1$ υπάρχουν πολλά σημεία εκτός των ορίων.

Στη φάση ΙΙ της λειτουργίας των $\bar{x}$ και R διαγραμμάτων ελέγχου σταθεροποιούνται τα αξιόπιστα όρια ελέγχου και τα διαγράμματα χρησιμοποιούνται για παρακολούθηση της μελλοντικής παραγωγής.

Το $\bar{x}$ διάγραμμα ελέγχου χρησιμοποιείται για να ανιχνεύσει μεγάλες αλλαγές σε μικρά δείγματα ($n = 5$) ή μικρές αλλαγές σε μεγάλα δείγματα ($n = 25$).

Το R διάγραμμα ελέγχου δεν είναι τόσο ευαίσθητο στις αλλαγές στην τυπική απόκλιση της διεργασίας για μικρά δείγματα. Για μεγάλα δείγματα ($n > 10$) συνίσταται η χρήση των S ή S^2 διαγραμμάτων ελέγχου.

1.4.3 $\bar{x}$ και S διαγράμματα ελέγχου

Τα $\bar{x}$ και S διαγράμματα ελέγχου χρησιμοποιούνται όταν το μέγεθος του δείγματος είναι μεγάλο (μεγαλύτερο του 10) ή δεν είναι σταθερό. Η κατασκευή τους είναι ίδια με αυτή των $\bar{x}$ και R με τη διαφορά ότι, για κάθε δείγμα πρέπει να υπολογίζουμε τα $\bar{x}$ και S. Εάν το σ^2 είναι άγνωστο τότε το εκτιμούμε από το $\hat{S}^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ με $\mu_S = E(S) = \sigma c_4$ και $\sigma_S = \sqrt{Var(S)} = \sigma\sqrt{1-c_4^2}$, όπου c_4 σταθερά η οποία εξαρτάται από το μέγεθος του δείγματος.

Τα όρια ελέγχου στο S διάγραμμα στη φάση II είναι:

$$\begin{aligned} UCL &= \left(c_4 + 3\sqrt{1-c_4^2}\right)\sigma \\ CL &= c_4\sigma \\ LCL &= \left(c_4 - 3\sqrt{1-c_4^2}\right)\sigma. \end{aligned}$$

Αν το σ είναι άγνωστο το εκτιμούμε από το $\hat{\sigma} = \bar{S}/c_4$.

Τα όρια ελέγχου του x διαγράμματος στη φάση I είναι:

$$\begin{aligned} UCL &= \bar{\bar{x}} + \frac{3}{c_4\sqrt{n}}\bar{S} \\ CL &= \bar{\bar{x}} \\ LCL &= \bar{\bar{x}} - \frac{3}{c_4\sqrt{n}}\bar{S}. \end{aligned}$$

Τα όρια ελέγχου του S διαγράμματος στη φάση I είναι:

$$\begin{aligned} UCL &= \left(1 + (3/c_4)\sqrt{1-c_4^2}\right)\bar{S} \\ CL &= \bar{S} \\ LCL &= \left(1 - (3/c_4)\sqrt{1-c_4^2}\right).\bar{S} \end{aligned}$$

Όταν το μέγεθος του δείγματος δεν είναι σταθερό τότε $\bar{\bar{x}} = \dfrac{\sum_{i=1}^{m} n_i\bar{x}_i}{\sum_{i=1}^{m} n_i}$ και $\bar{S} = \sqrt{\dfrac{\sum_{i=1}^{m}(n_i-1)S_i^2}{\sum_{i=1}^{m} n_i - m}}$. Αν τα n_i δεν είναι πολύ διαφορετικά μεταξύ τους μπορούμε να τα αντικαταστήσουμε με το $\bar{n}$ ($average\ sample\ size$).

1.4.4 Το S^2 διάγραμμα ελέγχου

Τα όρια ελέγχου S^2 διαγράμματος στη φάση I είναι:

$$\begin{aligned} UCL &= \frac{\bar{S}^2}{n-1}\chi^2_{\alpha/2,n-1} \\ CL &= \bar{S}^2 \\ LCL &= \frac{\bar{S}^2}{n-1}\chi^2_{1-\alpha/2,n-1}. \end{aligned}$$

1.4.5 Διαγράμματα Ελέγχου για μεμονωμένες παρατηρήσεις

Όταν $n = 1$ χρησιμοποιούνται τα διαγράμματα ελέγχου για μεμονωμένες παρατηρήσεις. Έστω $x \sim N(\mu, \sigma^2)$ με μ, σ γνωστά.

Τα όρια ελέγχου του x διαγράμματος είναι :

$$\begin{aligned} UCL_x &= \bar{x} + 3\sigma \\ CL_x &= \bar{x} \\ LCL_x &= \bar{x} - 3\sigma. \end{aligned}$$

Στα διαγράμματα ελέγχου για μεμονωμένες παρατηρήσεις χρησιμοποιείται το κινούμενο εύρος (*moving range* - MR) δύο παρατηρήσεων το οποίο ορίζεται ως εξής:

$$MR_i = \mid x_i - x_{i-1} \mid = max\{x_{i-1}, x_i\} - min\{x_{i-1}, x_i\}, i \geq 2. \tag{1.5}$$

Το MR διάγραμμα μπορεί να κατασκευαστεί ως εξής:

Τα όρια ελέγχου του MR διαγράμματος είναι:

$$\begin{aligned} UCL_{MR} &= D_2\sigma \\ CL_{MR} &= d_2\sigma \\ LCL_{MR} &= D_1\sigma. \end{aligned}$$

Αν μ, σ άγνωστα τα εκτιμούμε από τις τιμές:

$$\hat{\mu} = \bar{x} = \left(\sum_{i=1}^{m} x_i/m\right) \tag{1.6}$$

$$\hat{\sigma} = \overline{MR}/d_2. \tag{1.7}$$

όπου $\overline{MR} = (MR_1 + \ldots + MR_{i-1}) / (m-1)$. Τα D_1 και D_2 δίνονται στον Πίνακα 1.1.

1.5 Αθροιστικά διαγράμματα ελέγχου *(Cumulative Sum Control Chart, CUSUM)*

1.5.1 Εισαγωγή

Τα αθροιστικά διαγράμματα ελέγχου εισήχθησαν το 1954 από τον Βρετανό Page. Το βασικό τους πλεονέκτημα είναι ότι είναι ευαίσθητα στην εύρεση μικρών αλλαγών στη διεργασία και γι' αυτό χρησιμοποιούνται στην ανίχνευση μικρών συστηματικών σφαλμάτων. Σε αντίθεση με τα διαγράμματα ελέγχου *Shewhart*, τα αθροιστικά διαγράμματα ελέγχου είναι αποδοτικά στις περιπτώσεις μεμονωμένων παρατηρήσεων και ενσωματώνουν άμεσα όλη την πληροφορία από τις τιμές του δείγματος σχεδιάζοντας τα συσσωρευμένα αθροίσματα των αποκλίσεων των τιμών από την τιμή-στόχο μ_0. Έστω x το χαρακτηριστικό που μας ενδιαφέρει και $x \sim N(\mu_0, \sigma^2)$. Επιλέγοντας m τυχαία δείγματα x_i μεγέθους $n \geq 1$ το καθένα ο δειγματικός μέσος $\bar{x} \sim N(\mu_{\bar{x}}, \sigma^2_{\bar{x}})$, όπου $\mu_{\bar{x}}$=μ_0 και $\sigma^2_{\bar{x}}$=σ^2/n.

1.5.2 Διάγραμμα *Tabular Cusum*

Είναι ο προτιμότερος τρόπος αναπαράστασης των *cusum*. Κατασκευάζονται μέσω του υπολογισμού δύο συσσωρευμένων αθροισμάτων για κάθε τιμή ελέγχου. Οι θετικές αποκλίσεις από το στόχο αθροίζονται με το άνω συσσωρευμένο άθροισμα C_i^+ (*one sided upper cusum*), ενώ οι αρνητικές αποκλίσεις από το στόχο με το κάτω συσσωρευμένο άθροισμα C_i^- (*one sided lower cusum*). Αυτά τα αθροίσματα υπολογίζονται από τους τύπους:

$$C_i^+ = max[0, x_i - (\mu_0 + K) + C_{i-1}^+]$$
$$C_i^- = max[0, (\mu_0 + K) - x_i + C_{i-1}^-],$$

όπου $1 \leq i \leq m$, $C_0^+ = C_0^- = 0$ και $K = \dfrac{|\mu_1 - \mu_0|}{2}$ η τιμή αναφοράς (*reference value*).
Τα αθροίσματα C_i^+ και C_i^- υπολογίζονται απο τις τιμές των x_i από τη μέση τιμή μ_0, όταν αυτές είναι μεγαλύτερες από την τιμή αναφοράς K. Κάθε φορά που οι διαφορές $x_i - (\mu_0 + K)$, $(\mu_0 + K) - x_i$ γίνονται αρνητικές τα αθροίσματα μηδενίζονται και αυξάνονται ξανά όταν αυτές οι διαφορές γίνουν μεγαλύτερες του 0.
Η γραφική αναπαράσταση του *tabular cusum* γίνεται σχεδιάζοντας τα C_i^+ και C_i^- ως διαφορετικές στήλες πάνω και κάτω από τη μέση τιμή. Το όριο ελέγχου που σχεδιάζεται στο διάγραμμα είναι το διάστημα απόφασης H (*decision*

$interval$), και αναπαριστάται με 2 ευθείες, την H^+ και H^-, παράλληλες προς το μέσο μ_0. Οι ευθείες αυτές είναι τα επιτρεπτά όρια των αθροισμάτων C_i^+ και C_i^-. Είναι πολύ σημαντικός ο καθορισμός του H και του K. Συνήθως επιλέγουμε $H = h\sigma$ και $K = k\sigma$, όπου k είναι το μέγεθος της μετατόπισης που θέλουμε να ανιχνευθεί.

1.5.3 Τυποποιημένο *(standardized)* διάγραμμα *cusum*

Σε μερικές περιπτώσεις είναι προτιμότερο η μεταβλητή x_i να είναι τυποποιημένη πριν τον υπολογισμό των συσσωρευμένων αθροισμάτων. Έτσι ορίζεται η μεταβλητή

$$Y_i = \frac{x_i - \mu_0}{\sigma} \sim N(0,1)$$

η οποία είναι η τυποποιημένη τιμή της x_i. Τότε τα άνω και κάτω συσσωρευμένα αθροίσματα μετασχηματίζονται ως εξής:

$$C_i^+ = max[0, Y_i - K + C_{i-1}^+]$$
$$C_i^- = max[0, -K - Y_i + C_{i-1}^-],$$

όπου $1 \leq i \leq m$, $C_0^+ = C_0^- = 0$.
Τα τυποποιημένα διαγράμματα $cusum$ έχουν δύο πλεονεκτήματα:

1. Υπάρχουν πολλά διαγράμματα $cusum$ με τις ίδιες τιμές των k και h, αφού οι παράμετροι δεν εξαρτώνται από την τυπική απόκλιση κάθε διεργασίας.

2. Με τη χρήση της τυποποιημένης μεταβλητής Y_i δημιουργούνται εύκολα τα διαγράμματα για τον έλεγχο της μεταβλητότητας μίας διεργασίας.

1.5.4 Διαγράμματα *scale cusum*

Τα $scale\ cusum$ διαγράμματα χρησιμοποιούνται για την παρακολούθηση της μεταβλητότητας της διεργασίας. Έστω $x_i \sim N(\mu_0, \sigma^2)$, τότε η τυποποιημένη μεταβλητή είναι η $Y_i = \frac{x_i - \mu_0}{\sigma}$. Ο Hawkins (1981,1993) συνιστά τη δημιουργία μιας καινούριας τυποποιημένης μεταβλητής, της

$$v_i = \frac{\sqrt{|Y_i|} - 0.822}{0.349},$$

όπου $1 \leq i \leq m$. Η καινούρια αυτή μεταβλητή έχει ευαισθησία στις αλλαγές του μέσου και της διασποράς και ακολουθεί την $N(0,1)$.
Τα $scale\ cusum$ είναι:

$$S_i^+ = max[0, v_i - K + S_{i-1}^+]$$
$$S_i^- = max[0, -K - v_i + S_{i-1}^-],$$

όπου $S_0^+ = S_0^- = 0$ και τα K και H επιλέγονται όπως και στα διαγράμματα του μέσου.

Αν η τυπική απόκλιση της διεργασίας αυξάνεται, αυξάνονται και τα S_i^+ μέχρι να ξεπεράσουν το διάστημα απόφασης H και αν η τυπική απόκλιση μειώνεται, μειώνονται και τα S_i^- μέχρι να ξεπεράσουν το H. Αν στο διάγραμμα *scale cusum* υπάρχει εκτός ελέγχου ένδειξη, τότε υπάρχει υποψία μετατόπισης της διασποράς, ενώ αν υπάρχουν εκτός ελέγχου ενδείξεις και στο διάγραμμα του μέσου και στο διάγραμμα της μεταβλητότητας, τότε υπάρχει υποψία μετατόπισης του μέσου.

1.6 Διαγράμματα ελέγχου με κινητούς μέσους και εκθετικά βάρη (*Exponentially Weighted Moving Average, EWMA)*

1.6.1 Εισαγωγή

Το $EWMA$ (*Exponentially weighted moving average*) διάγραμμα ελέγχου παρουσιάστηκε από τον Roberts το 1959 και είναι ένα εναλλακτικό διάγραμμα των *Shewhart* διαγραμμάτων. Η χρήση του προτείνεται όταν θέλουμε να εντοπίσουμε μικρές μεταβολές στο μέσο μιας διεργασίας και όταν έχουμε μεμονωμένες παρατηρήσεις.

Ο εκθετικά κατανεμημένος κινητός μέσος ορίζεται από τη σχέση:

$$z_i{=}\lambda x_i{+}(1\text{-}\lambda)z_{i-1}$$

όπου x_i είναι οι παρατηρήσεις, $z_0 = \mu_0$ και $\lambda{\in}(0,1]$ είναι μία σταθερά, η οποία καλείται συντελεστής βαρύτητας ($weighting\ factor$) και καθορίζει το βαθμό κατά τον οποίο παλαιότερα δεδομένα εισάγονται στον υπολογισμό του $EWMA$.

1.6.2 Σχεδιασμός του $EWMA$

Η κατασκευή του $EWMA$ διαγράμματος απαιτεί τον υπολογισμό της μέσης τιμής και της τυπικής απόκλισης της z_i. Αν οι παρατηρήσεις x_i είναι ανεξάρτητες τυχαίες μεταβλητές με διασπορά σ^2, τότε η διασπορά του z_i είναι

$$\sigma_{z_i}^2{=}\sigma^2[\lambda/(2-\lambda)](1-(1-\lambda)^{2i})$$

Η κεντρική γραμμή και τα όρια ελέγχου του $EWMA$ διαγράμματος είναι:

$$\begin{aligned} UCL &= \mu_0 + L\sigma\sqrt{[\lambda/(2-\lambda)](1-(1-\lambda)^{2i})} \\ CL &= \mu_0 \\ LCL &= \mu_0 - L\sigma\sqrt{[\lambda/(2-\lambda)](1-(1-\lambda)^{2i})}, \end{aligned}$$

όπου L το εύρος των ορίων και μ_0 η τιμή-στόχος. Καθώς το i αυξάνεται, η τιμή $1-(1-\lambda)^{2i}$ τείνει στο 1, άρα από ένα σημείο και πέρα τα όρια ελέγχου μετατρέπονται σε

$$\begin{aligned} UCL &= \mu_0 + L\sigma\sqrt{[\lambda/(2-\lambda)]} \\ CL &= \mu_0 \\ LCL &= \mu_0 - L\sigma\sqrt{[\lambda/(2-\lambda)]}. \end{aligned}$$

Τότε τα όρια ελέγχου απεικονίζονται στο διάγραμμα σαν δύο ευθείες παράλληλες μεταξύ τους. Τα παραπάνω ισχύουν και για μεμονωμένες παρατηρήσεις. Επίσης είναι απαραίτητος ο προσδιορισμός των λ και L. Συνήθως χρησιμοποιείται η τιμή $L = 3$ και $0.05 \leq \lambda \leq 0.25$ έτσι ώστε να έχουμε καλή απόδοση. Υπάρχει ενδιαφέρον στη μελέτη των $EWMA$ με μικρο λ. Όταν οι τιμές του $EWMA$ βρίσκονται από τη μία πλευρά της κεντρικής γραμμής και προκύπτει αλλαγή σε μία τιμή από την άλλη πλευρά, το $EWMA$ αργεί λίγο να αντιδράσει στην αλλαγή εξαιτίας του λ. Αυτό καλείται φαινόμενο αδράνειας ($inertia\ effect$) και μειώνει την αποτελεσματικότητα του $EWMA$ σε μικρές αλλαγές. Το σήμα αντίδρασης ενός διαγράμματος ελέγχου ($signal\ resistance\ of\ a\ control\ chart$) είναι η μεγαλύτερη τυποποιημένη απόκλιση του μέσου από την τιμή στόχο και για το $Shewhart$ διάγραμμα ελέγχου ορίζεται ως

$$SR(\bar{x}) = L \tag{1.8}$$

και για το $EWMA$ ως

$$SR(EWMA) = \frac{L\sqrt{\frac{\lambda}{2-\lambda}} - (1-\lambda)w}{\lambda} \tag{1.9}$$

όπου w είναι η τιμή της $EWMA$ στατιστικής συνάρτησης. Η μέγιστη τιμή που μπορεί να πάρει το $SR(EWMA)$ είναι $L\sqrt{(2-\lambda)/\lambda}$, αν το διάγραμμα έχει ασυμπτωτικά όρια.

Το $EWMA$ συμπεριφέρεται καλά σε μικρές αλλαγές αλλά η απόδοσή του δεν είναι τόσο καλή όσο του διαγράμματος $Shewhart$, όταν οι αλλαγές στη διεργασία είναι μεγάλες. Ένας καλός τρόπος να αντιμετωπιστεί αυτό το πρόβλημα είναι η χρήση ενός συνδυασμένου $Shewhart$ διαγράμματος με ένα $EWMA$. Επίσης μπορούμε στο ίδιο διάγραμμα να σχεδιάσουμε το στατιστικό $\bar{x}$ και το στατιστικό του $EWMA$, z_i, μαζί με τα όρια και των δύο διαγραμμάτων.

1.7 Πολυμεταβλητός έλεγχος και παρακολούθηση διεργασιών

1.7.1 Εισαγωγή

Στην πραγματικότητα οι περισσότερες διεργασίες παρακολούθησης και ελέγχου περιέχουν περισσότερες από μία μεταβλητές. Η εφαρμογή του μεταβλητού ελέγχου διεργασίας για κάθε μία μεταβλητή ξεχωριστά δεν είναι αποτελεσματική και οδηγεί σε λανθασμένα συμπεράσματα. Στις περισσότερες περιπτώσεις χρειάζεται ταυτόχρονη παρακολούθηση ή έλεγχος δύο ή περισσότερων συσχετισμένων ποιοτικών χαρακτηριστικών. Γενικά αν έχουμε p στατιστικά ανεξάρτητα ποιοτικά χαρακτηριστικά και ένα $\bar{x}$ διάγραμμα ελέγχου με $P[$σφάλμα τύπου I$]=\alpha$ για κάθε χαρακτηριστικό, τότε η πιθανότητα σφάλματος τύπου I για το κοινό έλεγχο διεργασιών είναι $\alpha' = 1-(1-\alpha)^p$. Τα προβλήματα ελέγχου και παρακολούθησης στα οποία μας ενδιαφέρουν συσχετισμένες μεταβλητές καλούνται πολυμεταβλητά προβλήματα ελέγχου - ποιότητας ($multivariate\ quality-control\ problems$).

1.7.2 Περιγραφή των πολυμεταβλητών δεδομένων

Στο μονομεταβλητό στατιστικό έλεγχο ποιότητας, χρησιμοποιούμε την κανονική κατανομή για να περιγράψουμε τη συμπεριφορά ενός συνεχούς ποιοτικού χαρακτηριστικού, του οποίου η συνάρτηση πυκνότητας πιθανότητας είναι

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad -\infty < x < \infty, \tag{1.10}$$

όπου μ ο μέσος της κανονικής κατανομής και σ η τυπική απόκλιση. Η παραπάνω εξίσωση γράφεται και ως εξής:

$$(x-\mu)(\sigma^2)^{-1}(x-\mu) \tag{1.11}$$

Η ίδια προσέγγιση μπορεί να χρησιμοποιηθεί και για την περίπτωση της πολυμεταβλητής κανονικής κατανομής. Έστω ότι έχουμε p μεταβλητές $x_1, ...x_p$, τις οποίες τοποθετούμε σε ένα διάνυσμα, $\mathbf{x'} = [x_1, ...x_p]$. Έστω $\boldsymbol{\mu}'$=$[\mu_1, \mu_2, ..., \mu_p]$ είναι το διάνυσμα των μέσων και $\boldsymbol{\Sigma}$ είναι ένας $p \times p$ πίνακας συνδιασποράς, στον οποίο στην κύρια διαγώνιο βρίσκονται οι διασπορές των x και τα υπόλοιπα στοιχεία είναι οι συνδιασπορές.

Η σταθμισμένη τυποποιημένη απόσταση από το $\mathbf{x}$ στο $\boldsymbol{\mu}$ είναι:

$$\mathbf{(x\text{-}\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(x\text{-}\boldsymbol{\mu})} \tag{1.12}$$

Η πολυμεταβλητή συνάρτηση πυκνότητας πιθανότητας είναι

$$f(x) = \frac{1}{(2\pi)^{p/2} \mid \boldsymbol{\Sigma} \mid^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \qquad -\infty < \mathbf{x}_j < \infty, j = 1, 2, ..., p. \tag{1.13}$$

Όταν το δείγμα προέρχεται από μία πολυμεταβλητή κανονική κατανομή, το διάνυσμα του μέσου και ο πίνακας συνδιασποράς υπολογίζονται ως εξής:

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \tag{1.14}$$

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'. \tag{1.15}$$

Οι τιμές της κύριας διαγωνίου του πίνακα συνδιασποράς είναι:

$$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \tag{1.16}$$

και οι υπόλοιπες τιμές του πίνακα είναι:

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \tag{1.17}$$

1.7.3 Το διάγραμμα ελέγχου *Hotelling* T^2

Το *Hotelling* T^2 διάγραμμα ελέγχου είναι ένα διάγραμμα ελέγχου που χρησιμοποιείται πιο συχνά για την παρακολούθηση του διανύσματος των μέσων της διεργασίας και είναι ανάλογο με το μονομεταβλητό *Shewhart* $\bar{x}$ διάγραμμα ελέγχου.

1.7.3.1 Δεδομένα σε υποομάδες

Έστω ότι έχουμε δύο ποιοτικά χαρακτηριστικά x_1 και x_2, τα οποία ακολουθούν τη διδιάστατη κανονική κατανομή. Έστω μ_1 και μ_2 είναι οι μέσες τιμές τους, σ_1 και σ_2 οι τυπικές αποκλίσεις και σ_{12} η συνδιασπορά τους. Αν οι σ_1, σ_2, σ_{12} είναι γνωστές και $\bar{x_1}$ και $\bar{x_2}$ είναι οι δειγματικοί μέσοι των δύο ποιοτικών χαρακτηριστικών, τότε η στατιστική συνάρτηση:

$$\chi_0^2 = \frac{n}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}\left[\sigma_2^2\left(\bar{x}_1 - \mu_1\right)^2 + \sigma_1^2\left(\bar{x}_2 - \mu_2\right)^2 - 2\sigma_{12}\left(\bar{x}_1 - \mu_1\right)\left(\bar{x} - \mu_2\right)\right] \tag{1.18}$$

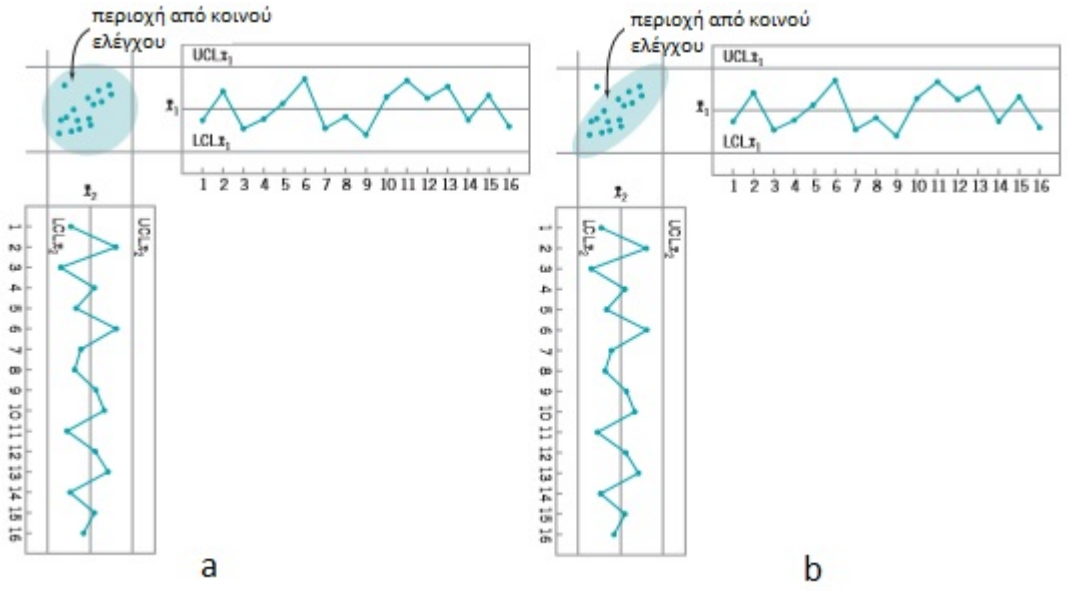
ακολουθεί την χ^2 με δύο βαθμούς ελευθερίας. Αυτή η εξίσωση χρησιμοποιείται σαν βάση του διαγράμματος ελέγχου για τους μέσους μ_1, μ_2 της διεργασίας. Αν η τιμή του χ_0^2 ξεπερνάει το άνω όριο, $UCL{=}\chi_{\alpha,2}^2$, τότε ο μέσος έχει αλλάξει σε μία εκτός-ελέγχου τιμή.

Η διαδικασία παρακολούθησης της διεργασίας παρουσιάζεται γραφικά. Έστω ότι οι μεταβλητές είναι ανεξάρτητες, οπότε $\sigma_{12} = 0$. Τότε η εξίσωση (1.18) ορίζει μία έλλειψη με κέντρο στα (μ_1, μ_2) με κύριους άξονες παράλληλους στα $\bar{x_1}$, $\bar{x_2}$, όπως φαίνεται στο Σχήμα 1.4*a*, το οποίο καλείται έλλειψη ελέγχου (*control ellipse*). Στην περίπτωση που οι μεταβλητές είναι εξαρτημένες και $\sigma_{12} \neq 0$ η αντίστοιχη έλλειψη ελέγχου φαίνεται στο Σχήμα 1.4*b*. Όταν οι δύο μεταβλητές είναι ανεξάρτητες, οι κύριοι άξονες δεν είναι πλέον παράλληλοι στους άξονες $\bar{x_1}$, $\bar{x_2}$.

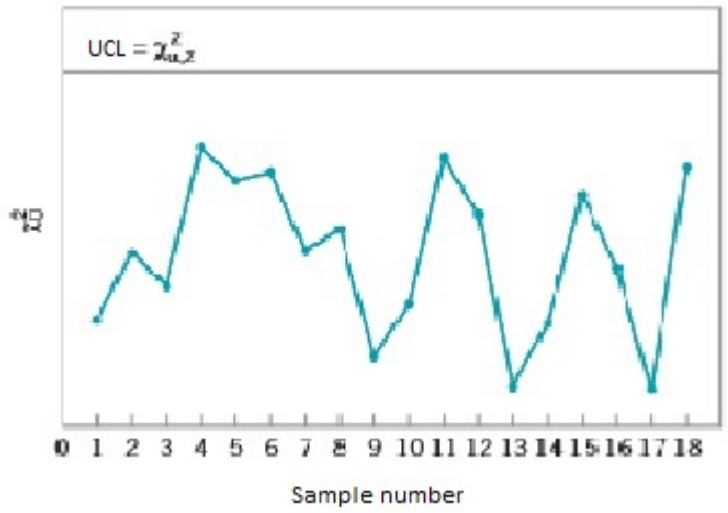
Αν τα σημεία σχεδιαστούν εντός της έλλειψης τότε η διεργασία είναι εντός ελέγχου. Ωστόσο υπάρχουν μειονεκτήματα που σχετίζονται με τον έλεγχο της έλλειψης. Πρώτον χάνεται η χρονική σειρά με την οποία εισήχθησαν τα δεδομένα, κάτι το οποίο μπορεί να ξεπεραστεί αριθμώντας τα δεδομένα και δεύτερον είναι δύσκολο να σχεδιαστεί η έλλειψη για περισσότερα από δύο ποιοτικά χαρακτηριστικά. Για να αποφευχθούν αυτά τα προβλήματα, συνήθως σχεδιάζονται οι τιμές του χ_0^2, υπολογισμένες από την παραπάνω εξίσωση, από κάθε δείγμα σε ένα διάγραμμα ελέγχου με μόνο το UCL στο $\chi_{\alpha,2}^2$, όπως φαίνεται στο Σχήμα 1.5. Αυτό το διάγραμμα καλείται χ^2 διάγραμμα ελέγχου ($chi-square\ control\ chart$).

Στα χ^2 διαγράμματα ελέγχου διατηρείται η σειρά εισαγωγής των σημείων και η κατάσταση της διεργασίας χαρακτηρίζεται μόνο από την τιμή χ_0^2, γι'αυτό και είναι εύχρηστα στην περίπτωση δύο ή περισσότερων ποιοτικών χαρακτηριστικών. Η επέκταση αυτών των αποτελεσμάτων στην περίπτωση p ποιοτικών χαρακτηριστικών γίνεται ως εξής:

Έστω ότι σε ένα $p{\times}1$ διάνυσμα έχουμε τους μέσους των ποιοτικών χαρακτηριστικών $\bar{\mathbf{x}}{=}[\bar{x_1}\ \bar{x_2}\ ...\ \bar{x_p}]^T$. Η στατιστική συνάρτηση ελέγχου που σχεδιάζεται



Σχήμα 1.4: Η έλλειψη ελέγχου για εξαρτημένες και ανεξάρτητες μεταβλητές.



Σχήμα 1.5: Ένα χ^2 διάγραμμα ελέγχου για p=2 ποιοτικά χαρακτηριστικά

στο χ^2 διάγραμμα ελέγχου είναι

$$\chi_0^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \tag{1.19}$$

και το άνω όριο του διαγράμματος ελέγχου είναι $UCL = \chi^2_{a,p}$.

Στις περισσότερες περιπτώσεις χρειάζεται να εκτιμήσουμε το $\boldsymbol{\mu}$ και το $\boldsymbol{\Sigma}$ από δείγματα που λαμβάνονται όταν η διεργασία είναι εντός ελέγχου. Έστω ότι έχουμε m τέτοια διαθέσιμα δείγματα. Τότε:

$$\bar{x}_{jk} = \frac{1}{n}\sum_{i=1}^{n} x_{ijk} \quad \begin{cases} j = 1, 2, \ldots, p \\ k = 1, 2, \ldots, m \end{cases}$$

$$s^2_{jk} = \frac{1}{n-1}\sum_{i=1}^{n} (x_{ijk} - \bar{x}_{jk})^2 \quad \begin{cases} j = 1, 2, \ldots, p \\ k = 1, 2, \ldots, m \end{cases},$$

όπου x_{ijk} είναι η i-οστή παρατήρηση του j-οστού ποιοτικού χαρακτηριστικού, στο k-οστό δείγμα. Η συνδιασπορά μεταξύ του ποιοτικού χαρακτηριστικού j και του ποιοτικού χαρακτηριστικού h στο k-οστό δείγμα είναι:

$$s_{jhk} = \frac{1}{n-1}\sum_{i=1}^{n} (x_{ijk} - \bar{x}_{jk})(x_{ihk} - \bar{x}_{hk}) \quad \begin{cases} j \neq h \\ k = 1, 2, \ldots, m \end{cases}.$$

Τότε οι μέσοι στα m δείγματα των $\bar{x}_{jk}, s^2_{jk}, s_{jhk}$ υπολογίζονται από τις σχέσεις:

$$\bar{\bar{x}}_j = \frac{1}{m}\sum_{k=1}^{m} \bar{x}_{jk} \quad j = 1, 2, \ldots, p \tag{1.20}$$

$$\bar{s}^2_j = \frac{1}{m}\sum_{k=1}^{m} s^2_{jk} \quad j = 1, 2, \ldots, p \tag{1.21}$$

$$\bar{s}_{jh} = \frac{1}{m}\sum_{k=1}^{m} s_{jhk} \quad j \neq h \tag{1.22}$$

και ο $p \times p$ μέσος του πίνακα συνδιασποράς:

$$\mathbf{S} = \begin{bmatrix} \bar{s}^2_1 & \bar{s}_{12} & \bar{s}_{13} & \ldots & \bar{s}_{1p} \\ & \bar{s}^2_2 & \bar{s}_{23} & \ldots & \bar{s}_{2p} \\ & & \bar{s}^2_3 & & \vdots \\ & & & \ddots & \bar{s}^2_p \end{bmatrix} \tag{1.23}$$

1.7.3.2 Το T^2 διάγραμμα ελέγχου

Έστω ότι ο παραπάνω πίνακας $\mathbf{S}$ εκτιμά το $\mathbf{\Sigma}$ και το διάνυσμα $\bar{\bar{x}}$ έχει υπολογιστεί από τις τιμές της διεργασίας όταν ήταν εντός ελέγχου. Η στατιστική συνάρτηση T^2 υπολογίζεται αν αντικαταστήσουμε στο στατιστικό του χ^2 το μ με $\bar{\bar{\mathbf{x}}}$ και το $\mathbf{\Sigma}$ με $\mathbf{S}$

$$T^2 = n\left(\bar{\mathbf{x}} - \bar{\bar{\mathbf{x}}}\right)' \mathbf{S}^{-1} \left(\bar{\mathbf{x}} - \bar{\bar{\mathbf{x}}}\right). \tag{1.24}$$

Το *Hotelling* T^2 είναι ένα κατευθυνόμενα αμετάβλητο διάγραμμα ελέγχου, του οποίου η ικανότητα να ανιχνεύει μετατοπίσεις στο διάνυσμα των μέσων εξαρτάται απ΄το μέγεθος της μετατόπισης και όχι από την κατευθυνσή της.

Τα όρια ελέγχου για το *Hotelling's* T^2 στατιστικό βασίζονται στον τρόπο χρήσης του διαγράμματος. Στη φάση Ι, όπου χρησιμοποιούμε το διάγραμμα ελέγχου για να καθορίσουμε τον έλεγχο έχουμε τα όρια:

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

ενώ στη φάση ΙΙ, όπου χρησιμοποιούμε το διάγραμμα ελέγχου για παρακολούθηση της μελλοντικής παραγωγής, τα όρια ελέγχου είναι:

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

Όταν το μέγεθος των δειγμάτων είναι πολύ μεγάλο, χρησιμοποιούμε και τα δύο άνω όρια με $UCL = \chi^2_{a,p}$ για αμφότερες τις φάσεις Ι και ΙΙ.

Όταν εμφανίζεται σήμα ότι η διεργασία είναι εκτός ελέγχου, στα πολυμεταβλητά διαγράμματα είναι δύσκολο να βρεθεί ποια από τις p μεταβλητές ευθύνεται για το σήμα. Σύνηθως γίνεται η γραφική παράσταση του μονομεταβλητού $\bar{x}$ διαγράμματος για κάθε $x_1, x_2, ..., x_p$ ξεχωριστά. Όμως αυτή η τακτική δεν είναι πάντα επιτυχημένη. Ο *Alt* (1985) πρότεινε τη χρήση των $\bar{x}$ διαγραμμάτων αντικαθιστώντας τα όρια ελέγχου $z_{a/2}$ με $z_{a/2p}$. Αυτή η προσέγγιση μειώνει τον αριθμό των λανθασμένων ειδοποιήσεων που συνδέονται με την ταυτόχρονη χρήση μονομεταβλητών διαγραμμάτων ελέγχου.

Μία άλλη χρήσιμη προσέγγιση της διάγνωσης ενός εκτός ελέγχου σήματος είναι η ανάλυση του T^2 στατιστικού σε συνιστώσες που αντανακλούν την συμβολή κάθε μεμονωμένης μεταβλητής. Αν T^2 είναι η τρέχουσα τιμή του στατιστικού και $T^2_{(i)}$ η τιμή του στατιστικού για όλες τις μεταβλητές εκτός από την i-οστή, τότε το $d_i = T^2 - T^2_{(i)}$ είναι ένας δείκτης μιας σχετικής συμβολής της i-οστής μεταβλητής σε όλο το στατιστικό. Οι μεγάλες τιμές του d_i δείχνουν ότι η διεργασία είναι εκτός ελέγχου λόγω της i μεταβλητής.

1.7.4 Μεμονωμένες παρατηρήσεις

Έστω ότι όλα τα m δείγματα έχουν μέγεθος $n = 1$ και p είναι ο αριθμός των ποιοτικών χαρακτηριστικών που παρατηρούνται σε κάθε δείγμα. Έστω $\bar{\mathbf{x}}$ και $\mathbf{S}$ το διάνυσμα των μέσων και ο πίνακας συνδιασποράς του δείγματος, αντίστοιχα. Το στατιστικό *Hotelling* T^2 ισούται με $T^2 = (\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}})$.

Τα όρια ελέγχου στη φάση II για το T^2 είναι:

$$UCL = \frac{p(m+1)(m-1)}{m^2 - mp} F_{\alpha,p,m-p}$$
$$LCL = 0$$

και όταν ο αριθμός των δειγμάτων m είναι πολύ μεγάλος ($m > 100$) χρησιμοποιούνται τα όρια ελέγχου:

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

ή

$$UCL = \chi^2_{\alpha,p}$$

Τα όρια ελέγχου στη φάση I είναι:

$$UCL = \frac{(m-1)^2}{m} \beta_{\alpha,p/2,(m-p-1)/2}$$
$$LCL = 0$$

όπου το $\beta_{\alpha,p/2,(m-p-1)/2}$ είναι το άνω α ποσοστιαίο σημείο μίας βήτα κατανομής με παραμέτρους $p/2$ και $(m-p-1)/2$.

Όταν ο πίνακας συνδιασποράς είναι άγνωστος μπορούμε να τον εκτιμήσουμε με διάφορους τρόπους. Ένας συνήθης εκτιμητής είναι ο

$$\mathbf{S}_1 = \frac{1}{m-1} \sum_{i=1}^{m} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' \tag{1.25}$$

Ένας δεύτερος εκτιμητής χρησιμοποιεί τη διαφορά μεταξύ πετυχημένων ζευγαριών των παρατηρήσεων $\mathbf{v}_i = \mathbf{x}_{i+1} - \mathbf{x}_i$, $i = 1, 2, ..., m-1$. Αρχικά προτάθηκε από τους *Holmes and Mergen* (1993) και φτιάχνει τον πίνακα V

$$\mathbf{V} = \begin{bmatrix} \mathbf{v}_1' \\ \mathbf{v}_2' \\ \vdots \\ \mathbf{v}_{m-1}' \end{bmatrix} \tag{1.26}$$

Τότε ο εκτιμητής του πίνακα συνδιασποράς $\mathbf{\Sigma}$ είναι: $\mathbf{S}_2 = \dfrac{1}{2} \dfrac{\mathbf{V}'\mathbf{V}}{(\mathrm{m}-1)}$.

1.7.5 Το πολυμεταβλητό *EWMA* διάγραμμα ελέγχου (*MEWMA*)

Τα διαγράμματα ελέγχου $Cusum$ και $EWMA$ διαφέρουν απ΄τα διαγράμματα τύπου $Shewhart$ γιατί παρέχουν μεγαλύτερη ευαισθησία σε μικρές αλλαγές στη μονομεταβλητή περίπτωση και μπορούν να επεκταθούν και στα πολυμεταβλητά προβλήματα ελέγχου ποιότητας. Όπως και στην μονομεταβλητή περίπτωση, αυτά τα διαγράμματα είναι στη φάση ΙΙ της διαδικασίας. Το $MEWMA$ (Lowry et al. (1992)) είναι λογική επέκταση του $EWMA$ και καθορίζεται από το

$$\mathbf{Z}_i = \lambda \mathbf{x}_i + (1-\lambda)\mathbf{Z}_{i-1} \tag{1.27}$$

όπου $0 \leq \lambda \leq 1$ και $\mathbf{Z}_0 = 0$. Η ποσότητα που σχεδιάζεται στο διάγραμμα ελέγχου είναι η $T_i^2 = \mathbf{Z}_i' \mathbf{\Sigma}_{\mathbf{Z}_i}^{-1} \mathbf{Z}_i$ και ο πίνακας συνδιασποράς είναι:

$$\mathbf{\Sigma}_{\mathbf{Z}_i} = \frac{\lambda}{2-\lambda}[1-(1-\lambda)^{2i}]\mathbf{\Sigma} \tag{1.28}$$

Για λ=1, το $MEWMA$ είναι ισοδύναμο με το T^2 διάγραμμα ελέγχου, αν και είναι πιο ευαίσθητο σε μικρές αλλαγές, όπως και στην μονομεταβλητή περίπτωση. Το μόνο που χρειαζόμαστε για να χαρακτηρίσουμε την απόδοση για κάθε μετατόπιση στο διάνυσμα των μέσων είναι αντίστοιχο της τιμής $\delta = (\boldsymbol{\mu}' \mathbf{\Sigma} \boldsymbol{\mu})^{1/2}$.

Το διάγραμμα ελέγχου $MEWMA$ είναι μία πολύ χρήσιμη διαδικασία, καθώς είναι σχετικά εύκολο να εφαρμοστεί και οι κανόνες σχεδιασμού του είναι καλά τεκμηριωμένοι.

1.7.6 Το πολυμεταβλητό *CUSUM* διάγραμμα ελέγχου (*MCUSUM*)

Το πολυμεταβλητό $CUSUM$ διάγραμμα ελέγχου χωρίζεται σε 2 κατηγορίες. Η πρώτη αφορά στην παρακολούθηση του διανύσματος του μέσου της διεργασίας και η δεύτερη στην παρακολούθηση του πίνακα διασποράς.

Έστω $\mathbf{X}_1$, $\mathbf{X}_2$,... μία σειρά παρατηρήσεων στη φάση ΙΙ μιας διεργασίας παραγωγής p-διαστάσεων οι οποίες ακολουθούν την $N_p(\mu_\mathbf{0}, \mathbf{\Sigma_0})$, όπου $\mu_\mathbf{0}, \mathbf{\Sigma_0}$ γνωστά.

Ο Woodall και ο Ncube (1985) πρότειναν την παρακολούθηση των p ανεξάρτητων συστατικών της διεργασίας με p μονομεταβλητά $CUSUM$ διαγράμματα. Το $MCUSUM$ δίνει σήμα ότι η διεργασία είναι εκτός ελέγχου, όταν ένα από τα μονομεταβλητά $CUSUM$ δώσει αντίστοιχο σήμα.

Χωρίς βλάβη της γενικότητας θεωρούμε ότι $\mu_\mathbf{0}$=0. Για τη j-οστή συνιστώσα $(1 \leq j \leq p)$ έχουμε ότι:

$$\begin{aligned} C_{n,j}^+ &= max\left(0, C_{n-1,j}^+ + X_{nj} - k_j\right) \qquad n \geq 1 \\ C_{n,j}^- &= min\left(0, C_{n-1,j}^- + X_{nj} + k_j\right), \end{aligned}$$

όπου $\mathbf{X}_n = (X_{n1}, X_{n2}, ..., X_{np})'$ είναι το n-οστό διάνυσμα παρατηρήσεων, $C^+_{0,j} = C^-_{0,j} = 0$ και k_j τιμή αναφοράς.

Αν $C^+_{n,j} > h_j$ ή $C^-_{n,j} < -h_j$, τότε το $CUSUM$ διάγραμμα δίνει ένα εκτός ελέγχου σήμα στη j-οστή συνιστώσα, στο n-οστό χρονικό σημείο. Το h_j είναι όριο ελέγχου.

Για την παρακολούθηση του πίνακα διασποράς αρκεί να αντικαταστήσουμε τα $C^+_{n,j}$, $C^-_{n,j}$ με τα στατιστικά για αναγνώριση μετατοπίσεων στη διασπορα. Τότε το $MCUSUM$ θα αναγνωρίζει τις αλλαγές στον πίνακα συνδιασποράς.

$$C^+_n = max\left(0, C^+_{n-1} + \left(\frac{X_n - \mu_0}{\sigma_0}\right)^2 - k\right) \qquad C^+_0 = 0$$

$$k = \frac{2log\left(\frac{\sigma_0}{\sigma_1}\right)}{\left(\frac{\sigma_0}{\sigma_1}\right)^2 - 1}$$

$$C^-_n = min\left(0, C^-_{n-1} + \left(\frac{X_n - \mu_0}{\sigma_0}\right)^2 - k\right) \qquad C^-_0 = 0.$$

Αν $C^+_n > h_u$ ή $C^-_n < h_L$ θα εμφανιστεί ένα εκτός ελέγχου σήμα.

1.7.7 Διαγράμματα ελέγχου για παρακολούθηση της μεταβλητότητας

Η παρακολούθηση πολυμεταβλητών διεργασιών χρειάζεται προσοχή σε δύο σημεία. Είναι σημαντικό να παρακολουθείται το διάνυσμα των μέσων $\boldsymbol{\mu}$ και η μεταβλητότητα της διεργασίας, η οποία δίνεται από τον $p \times p$ πίνακα συνδιασποράς $\boldsymbol{\Sigma}$.

Ο Alt (1985) παρουσίασε δύο χρήσιμες διαδικασίες. Η πρώτη διαδικασία είναι μία άμεση επέκταση του μονομεταβλητού S^2 διαγράμματος ελέγχου. Η διεργασία είναι ισοδύναμη με τους επαναλαμβανόμενους ελέγχους της σημαντικότητας της υπόθεσης ότι ο πίνακας διασποράς είναι ίσος με τον πίνακα των σταθερών $\boldsymbol{\Sigma}$. Το στατιστικό που σχεδιάζεται στο διάγραμμα ελέγχου για την i-στή παρατήρηση είναι

$$W_i = -pn + pn\ln(n) - nln\left(\mid \mathbf{A}_i \mid / \mid \boldsymbol{\Sigma} \mid\right) + tr\left(\boldsymbol{\Sigma}^{-1}\mathbf{A}_i\right), \qquad (1.29)$$

όπου $\mathbf{A}_i = (n-1)\mathbf{S}_i$. Αν το W_i σχεδιάζεται εκτός του ορίου $UCL = \chi^2_{\alpha,p(p+1)/2}$ η διεργασία είναι εκτός ελέγχου.

Η δεύτερη προσέγγιση βασίζεται στη γενικευμένη απόκλιση του δείγματος, $|\mathbf{S}|$. Αυτό το στατιστικό, το οποίο είναι η ορίζουσα του πίνακα διασποράς, χρησιμοποιείται ευρέως για τη μέτρηση πολυμεταβλητής διασποράς. Ισχύει ότι:

$$
\begin{aligned}
E(|\mathbf{S}|) &= b_1|\mathbf{\Sigma}| \\
V(|\mathbf{S}|) &= b_2|\mathbf{\Sigma}|^2 \\
b_1 &= \frac{1}{(n-1)^p}\prod_{i=1}^{p}(n-i) \\
b_2 &= \frac{1}{(n-1)^{2p}}\prod_{i=1}^{p}(n-i)\left[\prod_{j=1}^{p}(n-j+2) - \prod_{j=1}^{p}(n-j)\right].
\end{aligned}
$$

Τα όρια ελέγχου για το $|\mathbf{S}|$ είναι:

$$
\begin{aligned}
UCL &= |\mathbf{\Sigma}|(b_1 + 3b_2^{1/2}) \\
CL &= b_1|\mathbf{\Sigma}| \\
LCL &= |\mathbf{\Sigma}|(b_1 - 3b_2^{1/2}).
\end{aligned}
$$

Όταν το LCL παίρνει αρνητική τιμή, αντικαθιστάται με το 0.

Συνήθως το $\mathbf{\Sigma}$ εκτιμάται από τον πίνακα συνδιασπορών $\mathbf{S}$. Σε αυτή την περίπτωση στα όρια ελέγχου το $|\mathbf{\Sigma}|$ αντικαθιστάται με το $\mathbf{S}/b_1$.

2 Εισαγωγή στις μηχανές διανυσμάτων υποστήριξης

2.1 Εισαγωγή

Ο όρος Μηχανή Διανυσμάτων Υποστήριξης (*Support Vector Machine-SVM*) εμφανίστηκε το 1992 και εισήχθη από τους Boser, Guyon και Vapnik. Οι μηχανές διανυσμάτων υποστήριξης αποτελούν ένα εργαλείο για την επεξεργασία δεδομένων που χρησιμοποιείται σε πολλές εφαρμογές. Είναι ένα σύνολο συσχετισμένων μεθόδων εκμάθησης με επίβλεψη, που χρησιμοποιούνται στην ταξινόμηση και την παλινδρόμηση και ανήκουν σε μια οικογένεια γενικευμένων γραμμικών ταξινομητών. Επίσης μπορούν να χρησιμοποιηθούν για αναπαραστάσεις νευρωνικών δικτύων, *splines*, πολυώνυμων εκτιμητών κ.α. Η κεντρική ιδέα πίσω τα SVM μπορεί να συνοψιστεί στο εξής: **Δοθέντος ενός δείγματος (*train sample*) εκπαίδευσης η μηχανή διανυσμάτων υποστήριξης κατασκευάζει ένα υπερεπίπεδο ως επιφάνεια απόφασης με τρόπο τέτοιο ώστε το περιθώριο διαχωρισμού μεταξύ θετικών και αρνητικών παραδειγμάτων να μεγιστοποιείται.**

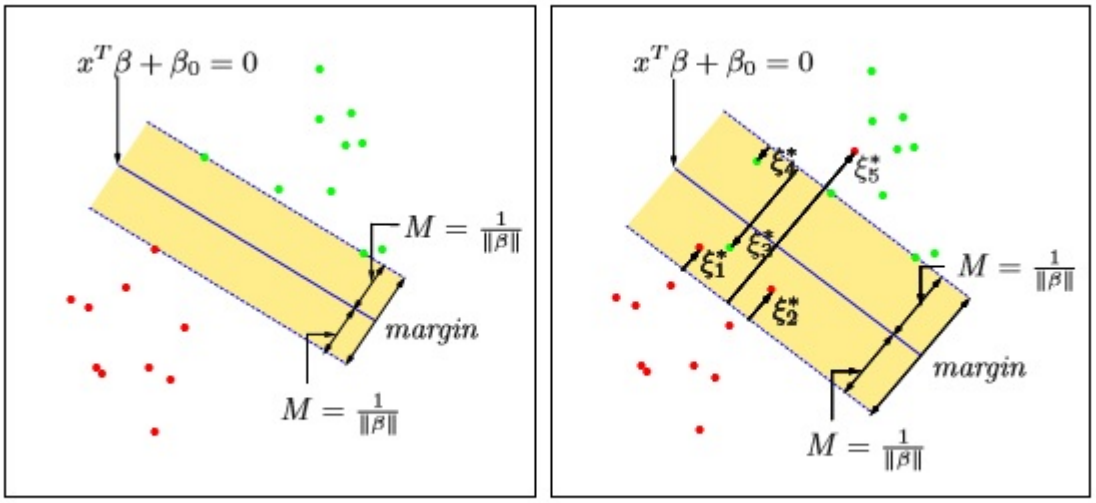
2.2 Ο ταξινομητής διανυσμάτων υποστήριξης

Η απλούστερη μορφή επίλυσης ενός προβλήματος είναι η δυαδική ταξινόμηση (*binary classification*), όπου πρέπει να γίνει ένας διαχωρισμός αντικειμένων που ανήκουν σε μία από τις δύο κατηγορίες. Η λογική μιας μηχανής εκμάθησης είναι να δίνει την τιμή y_i μιας συνάρτησης που αντιστοιχεί σε δοσμένο σημείο x_i. Αυτό γίνεται ως εξής:
Για δεδομένο σύνολο l σημείων x_i και έχοντας τις αντίστοιχες τιμές y_i που παίρνει η άγνωστη συνάρτηση, εκπαιδεύουμε τη μηχανή εκμάθησης να μάθει τη σχέση που συνδέει τα x_i με τα y_i. Το σύνολο των l σημείων αποτελείται από δύο υποσύνολα k, n. Έτσι το αποτέλεσμα της συνάρτησης θα είναι (+1) ή (-1), ανάλογα σε ποιο υποσύνολο ανήκει το δοθέν σημείο x_i. Τα δύο αυτά υποσύνολα ονομάζονται κλάσεις και τα ζεύγη (x_i, y_i) είναι τα δεδομένα εκπαίδευσης (*training data*).
Το υπερεπίπεδο καθορίζεται από το σύνολο

$$\{x : f(x) = x^T\beta + \beta_0 = 0\}, \tag{2.1}$$

όπου β είναι μοναδιαίο διάνυσμα ($\|\beta\|$=1). Αν οι κλάσεις είναι διαχωρίσιμες, μπορεί να βρεθεί εξίσωση $f(x) = x^T\beta + \beta_0$ με $y_i f(x_i) > 0 \ \forall \ i$, οπότε μπορεί

να βρεθεί το υπερεπίπεδο που δημιουργεί το μέγιστο περιθώριο ανάμεσα στα σημεία εκπαίδευσης και τις κλάσεις -1 και 1 (Σχήμα 2.1).



Σχήμα 2.1: Ένας ταξινομητής στη διαχωρίσιμη περίπτωση.

Το πρόβλημα βελτιστοποίησης

$$\max_{\beta,\beta_0,\|\beta\|=1} M, \text{ δεδομένου ότι } y_i(x_i^T\beta + \beta_0) \geq M, i = 1, 2, ..., N \qquad (2.2)$$

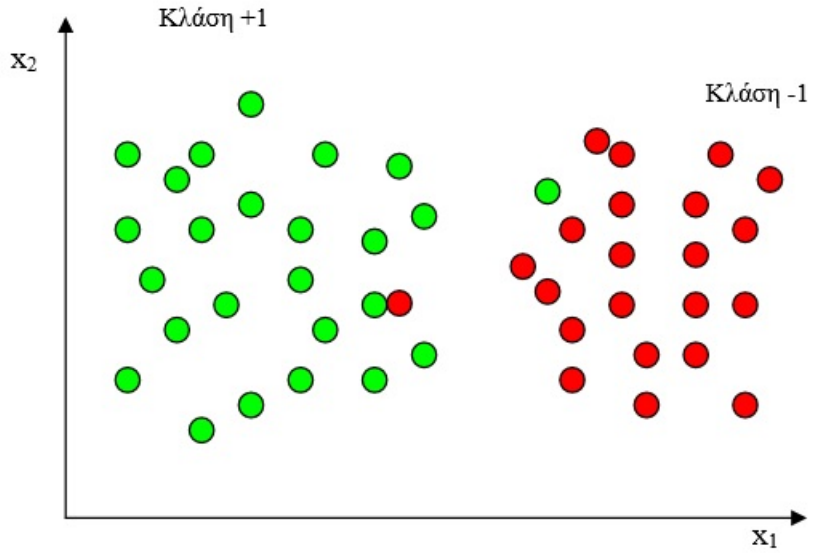
είναι κατάλληλο για τη δημιουργία του κατάλληλου περιθωρίου. Το M καλείται περιθώριο ($margin$) και ισχύει ότι M=$1/\|\beta\|$, οπότε το πρόβλημα βελτιστοποίησης είναι ισοδύναμο με το

$$\min_{\beta,\beta_0} \|\beta\|, \text{ δεδομένου ότι } y_i(x_i^T\beta + \beta_0) \geq 1, i = 1, 2, ..., N. \qquad (2.3)$$

Αυτός είναι ο πιο συχνός τρόπος γραφής του κριτηρίου διανυσμάτων υποστήριξης για δεδομένα που μπορούν να διαχωριστούν.

Στην περίπτωση που τα δεδομένα δεν είναι γραμμικώς διαχωρίσιμα, είναι αναγκαία η χαλάρωση των περιορισμών (2.3), έτσι ώστε να επιτρέπεται κάποια σημεία να μην είναι σωστά ταξινομημένα.
Αυτό γίνεται με την εισαγωγή μιας θετικής χαλαρής μεταβλητής, έστω ξ=(ξ_1,ξ_2,...,ξ_N). Υπάρχουν δύο τρόποι να τροποποιήσουμε τον περιορισμό (2.2):

$$y_i(x_i^T\beta + \beta_0) \geq M - \xi_i \qquad (2.4)$$



Σχήμα 2.2: Μη γραμμικά διαχωρίσιμα δεδομένα.

ή

$$y_i(x_i^T\beta + \beta_0) \geq M(1 - \xi_i) \tag{2.5}$$

$\forall i,\ \xi_i \geq 0,\ \sum_{i=1}^{N} \xi_i \leq$ *σταθερά*. Οι δύο αυτοί τρόποι οδηγούν σε διαφορετικά αποτελέσματα, αφού η πρώτη έχει ως αποτέλεσμα μη κυρτό πρόβλημα βελτιστοποίησης και η δεύτερη οδηγεί στο κλασικό ταξινομητή διανυσμάτων υποστήριξης, δηλαδή σε κυρτό πρόβλημα και αυτό θα μας απασχολήσει παρακάτω. Η τιμή ξ_i στον περιορισμό (2.5), είναι η αναλογική ποσότητα κατά την οποία η πρόβλεψη $f(x_i)$ βρίσκεται στη λανθασμένη πλευρά του ορίου. Έτσι φράσσοντας το $\sum_{i=1}^{N} \xi_i$, φράσσουμε την συνολικη αναλογική ποσότητα από την οποία οι προβλέψεις βρίσκονται στην λανθασμένη πλευρά των περιθωρίων. Οπότε προσπαθώντας να μειώσουμε τον αριθμό των μη ταξινομημένων σημείων, μετατρέπουμε το πρόβλημα ελαχιστοποίησης (2.3) ως εξής:

$$\min \|\beta\|, \text{ δεδομένου ότι } y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i, \forall i \text{ και } \xi_i \geq 0, \sum_{i=1}^{N} \xi_i \leq \textit{σταθερά} \tag{2.6}$$

Τα σημεία που βρίσκονται εντός των ορίων της κλάσης, δεν παίζουν μεγάλο ρόλο στο σχηματισμό των ορίων.

2.3 Υπολογισμός του ταξινομητή διανυσμάτων υποστήριξης

Το πρόβλημα (2.6) είναι τετραγωνικό με γραμμικούς περιορισμούς και μπορούμε να περιγράψουμε την τετραγωνική προγραμματιστική λύση του χρησιμοποιώντας τους πολλαπλασιαστές *Lagrange*. Παρακάτω δίνεται μία ισοδύναμη μορφή του προβλήματος (2.6).

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i \tag{2.7}$$

$$\text{δεδομένου ότι } y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i, \forall i \text{ και } \xi_i \geq 0, \tag{2.8}$$

όπου το C αντικαθιστά τη σταθερά που χρησιμοποιείται στο πρόβλημα (2.6). Η διαχωρίσιμη περίπτωση αντιστοιχεί σε $C = \infty$.
Η συνάρτηση *Lagrange* (πρωτογενής) είναι η εξής:

$$L_P = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i[y_i(x_i^T\beta + \beta_0) - (1 - \xi_i)] - \sum_{i=1}^{N}\mu_i\xi_i \tag{2.9}$$

και θέτοντας την παράγωγο ίση με 0, λαμβάνουμε:

$$\beta = \sum_{i=1}^{N}\alpha_i y_i x_i \tag{2.10}$$

$$0 = \sum_{i=1}^{N}\alpha_i y_i \tag{2.11}$$

$$\alpha_i = C - \mu_i, \forall i, \tag{2.12}$$

όπου $\alpha_i, \mu_i, \xi_i \geq 0\ \forall i$. Αντικαθιστώντας τις σχέσεις (2.10),(2.11),(2.12) στην σχέση (2.9), υπολογίζουμε τη διπλή *Lagrangian* αντικειμενική συνάρτηση:

$$L_D = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N}\alpha_i\alpha_{i'}y_i y_{i'}x_i^T x_{i'}, \tag{2.13}$$

η οποία δίνει ένα κάτω φράγμα για την αντικειμενική συνάρτηση (2.7) για κάθε εφικτό σημείο. Μεγιστοποιούμε την L_D δεδομένου ότι $0 \leq \alpha_i \leq C$ και $\sum_{i=1}^{N}\alpha_i y_i = 0$. Εκτός από τις σχέσεις (2.10),(2.11),(2.12) οι $Karush - Kuhn - Tucker$ συνθήκες περιλαμβάνουν τους περιορισμούς:

$$\alpha_i[y_i(x_i^T\beta + \beta_0) - (1 - \xi_i)] = 0 \tag{2.14}$$

$$\mu_i\xi_i = 0 \tag{2.15}$$

$$y_i(x_i^T\beta + \beta_0) - (1 - \xi_i) = 0. \tag{2.16}$$

Αυτές οι 6 εξισώσεις χαρακτηρίζουν μοναδικά τη λύση του αρχικού προβλήματος και του διπλού (2.13). Το β υπολογίζεται από τον τύπο:

$$\hat{\beta} = \sum_{i=1}^{N} \hat{\alpha}_i y_i x_i \tag{2.17}$$

όπου οι συντελεστές $\hat{\alpha}_i$ είναι διάφοροι του 0 μόνο για τις παρατηρήσεις i για τις οποίες οι περιορισμοί (2.14-2.16) ικανοποιούνται ακριβώς. Αυτές οι παρατηρήσεις καλούνται διανύσματα υποστήριξης (*support vectors*). Κάποια από αυτά τα σημεία υποστήριξης θα βρίσκονται πάνω στην ακμή του περιθωρίου και για αυτά θα ισχύει οτι $\hat{\xi}_i$=0 και θα έχουν $0 \leq \hat{\alpha}_i \leq C$ και τα υπόλοιπα σημεία θα έχουν $\hat{\xi}_i > 0$ και $\hat{\alpha}_i = C$.

Η μεγιστοποίηση του διπλού προβλήματος (2.13), είναι ένα πιο απλό πρόβλημα κυρτού τετραγωνικού προγραμματισμού από το αρχικό (2.9) και μπορεί να λυθεί με τις κλασικές τεχνικές (Murray et al. (1981)).

Δοθέντος των $\hat{\beta}_0$, $\hat{\beta}$, η συνάρτηση απόφασης μπορεί να γραφεί ως

$$\hat{G}(x) = sign[\hat{f}(x)] \tag{2.18}$$

$$= sign[x^T\hat{\beta} + \hat{\beta}_0]. \tag{2.19}$$

2.4 Μηχανές διανυσμάτων υποστήριξης και πυρήνες

Ο ταξινομητής διανυσμάτων υποστήριξης που περιγράφεται παραπάνω, βρίσκει γραμμικά όρια στο χώρο εισαγωγής (*input space*). Όπως και με άλλες γραμμικές μεθόδους, μπορούμε να κάνουμε τη μέθοδο πιο ευέλικτη μεγαλώνοντας το χώρο χαρακτηριστικών (*feature space*) χρησιμοποιώντας βάση επέκτασης. Γενικότερα τα γραμμικά όρια σε επεκταμένους χώρους έχουν καλύτερα αποτελέσματα στο διαχωρισμό των κλάσεων και μετατρέπονται σε μη-γραμμικά όρια στον αρχικό χώρο. Όταν επιλεχθούν οι συναρτήσεις βάσης, έστω $h_m(x)$, $m = 1, 2, ..., M$ η μέθοδος είναι ίδια με πριν. Προσαρμόζουμε τον ταξινομητή διανυσμάτων υποστήριξης, χρησιμοποιώντας τα χαρακτηριστικά εισόδου $h(x_i)$, $i = 1, 2, ..., N$ και παράγουμε την (μη-γραμμική) συνάρτηση $\hat{f}(x) = h(x)^T\hat{\beta} + \hat{\beta}_0$. Ο ταξινομητής είναι όπως πριν: $\hat{G}(x) = sign[\hat{f}(x)]$.

Ο ταξινομητής μηχανών διανυσμάτων υποστήριξης (*SVM classifier*) είναι μια επέκταση αυτής της ιδέας, όπου η διάσταση του επεκταμένου χώρου μπορεί να είναι πολύ μεγάλη, ακόμα και άπειρη σε μερικές περιπτώσεις.

Υπολογισμός του SVM για ταξινόμηση

Μπορούμε να παρουσιάσουμε τη αρχική συνάρτηση *Lagrange* (2.9)

$$L_P = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i[y_i(x_i^T\beta + \beta_0) - (1-\xi_i)] - \sum_{i=1}^{N}\mu_i\xi_i$$

και τη λύση της με άλλο τρόπο, ο οποίος περιέχει τα χαρακτηριστικά εισαγωγής μέσω εσωτερικών γινομένων. Αυτό γίνεται με τον μετασχηματισμό των χαρακτηριστικών διανυσμάτων $h(x_i)$. Για συγκεκριμένες τιμές του h, τα εσωτερικά γινόμενα μπορούν να υπολογιστούν πολύ εύκολα.
Η διπλή *Lagrange* συνάρτηση έχει τη μορφή:

$$L_D = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N}\alpha_i\alpha_{i'}y_iy_{i'}\langle h(x_i), h(x_{i'})\rangle. \tag{2.20}$$

Τότε η λύση της συνάρτησης $f(x)$ μπορεί να γραφεί:

$$f(x) = h(x)^T\beta + \beta_0 = \sum_{i=1}^{N}\alpha_iy_i\langle h(x), h(x_i)\rangle + \beta_0. \tag{2.21}$$

Όπως και πριν, το α_i και το β_0 μπορούν να υπολογιστούν λύνοντας την $y_if(x_i) = 1$ για κάθε x_i τέτοιο ώστε $0 < \alpha_i < C$.

Δεν χρειάζεται να καθορίσουμε το μετασχηματισμό της $h(x)$ εξ΄ ολοκλήρου, αρκεί να γνωρίζουμε τη συνάρτηση πυρήνα:

$$K(x, x') = \langle h(x), h(x')\rangle, \tag{2.22}$$

η οποία υπολογίζει το εσωτερικό γινόμενο των εικόνων που παράγονται στο χώρο χαρακτηριστικών βάσει της h δύο σημείων δεδομένων στο χώρο εισόδου. Η εφαρμογή της εξίσωσης (2.21) καλείται συνήθως τέχνασμα του πυρήνα. Το K πρέπει να είναι συμμετρική θετικά ορισμένη συνάρτηση. Υπάρχουν 3 δημοφιλείς επιλογές για το K στη βιβλιογραφία για τα SVM

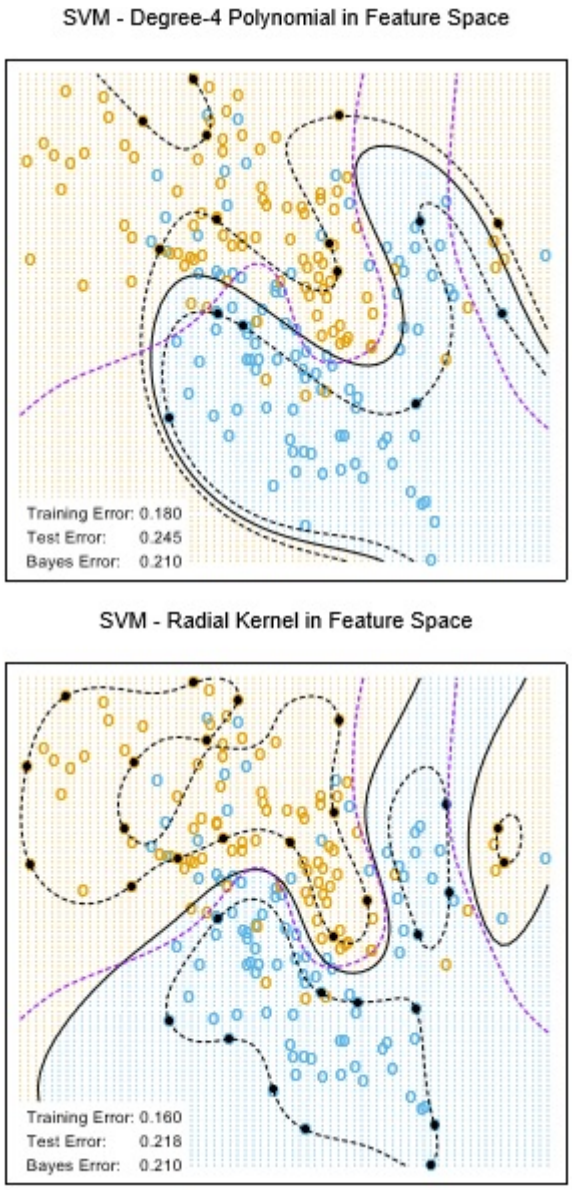
Πολυώνυμο βαθμού δ (dth-Degree polynomial) $: K(x, x') = (1 + \langle x, x'\rangle)^d$

Ακτινική βάση (Radial basis) $: K(x, x') = exp(-\gamma\|x - x'\|^2)$

Νευρωνικά δίκτυα (Neural network) $: K(x, x') = tanh(\kappa_1\langle x, x'\rangle) + \kappa_2)$

Άρα η λύση της συνάρτησης $f(x)$ γράφεται και ως:

$$f(x) = \sum_{i=1}^{N}\hat{\alpha}_iy_iK(x, x_i) + \hat{\beta}_0 \tag{2.23}$$



Σχήμα 2.3: (διακεκομμένη γραμμή) Δύο μη γραμμικά $SVMs$ που εφαρμόζονται σε μεικτά δεδομένα. Στο πρώτο γράφημα χρησιμοποιείται $4^{ου}$ βαθμού πολυωνυμικός πυρήνας και στο δεύτερο ακτινικής βάσης πυρήνας (με γ=1). Η παράμετρος κανονικοποίησης (C) έχει επιλεχθεί και στις δύο περιπτώσεις ώστε να πετυχαίνει καλό σφάλμα ελέγχου. Ο πυρήνας ακτινικής βάσης παράγει ένα όριο σχεδόν όμοιο με αυτό του $Bayes$, το οποίο είναι αυτό με τη μωβ γραμμή.

2.5 Το SVM ως μία ποινικοποιημένη μέθοδος

Με το $f(x) = h(x)^T\beta + \beta_0$, θεωρούμε το πρόβλημα βελτιστοποίησης

$$\min_{\beta_0,\beta} \sum_{i=1}^{N} [1 - y_i f(x_i)]_+ + \frac{\lambda}{2}\|\beta\|^2 \tag{2.24}$$

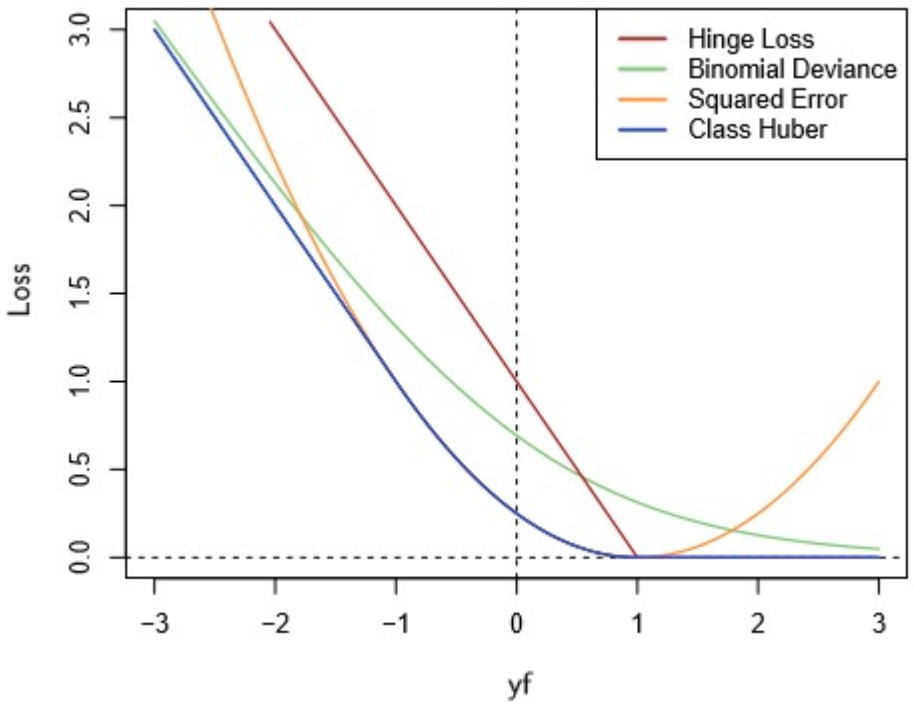
όπου ο δείκτης $''+''$ υποδηλώνει θετικό κομμάτι. Αυτό έχει τη μορφή απώλεια + ποινή, το οποίο είναι συνηθισμένο παράδειγμα στην εκτίμηση συναρτήσεων. Η λύση της (2.24) για $\lambda = 1/C$ είναι ίδια με την (2.7).

Εξετάζοντας την *hinge* συνάρτηση απώλειας $L(y, f) = [1-yf]_+$, βλέπουμε ότι είναι λογική για την ταξινόμηση δύο κλάσεων, όταν συγκρίνεται με άλλες συναρτήσεις απώλειας. Το Σχήμα 2.4 συγκρίνει την απώλεια λογαριθμικής πιθανοφάνειας για τη λογιστική παλινδρόμηση και την απώλεια του τετραγωνικού σφάλματος και μία παραλλαγή αυτών. Η αρνητική λογαριθμική πιθανοφάνεια ή διωνυμική απόκλιση έχει παρόμοια ουρά με την SVM απώλεια (*hing loss*), και δίνει μηδενική ποινή σε σημεία που είναι εντός του περιθωρίου και γραμμική ποινή στα σημεία που βρίσκονται στη λάθος πλευρά ή πολύ μακριά. Από την άλλη πλευρά, το σφάλμα των τετραγώνων δίνει τετραγωνική ποινή και σημεία εντός του περιθωρίου έχουν σημαντική επιρροή στο μοντέλο.
Η τετραγωνική *hinge* απώλεια $L(y, f) = [1 - yf]_+^2$ είναι όπως και η τετραγωνική, εκτός από το ότι είναι 0 για τα σημεία εντός του περιθωρίου. Οι Rosset και Zhu (2007) πρότειναν μία *huberized* εκδοχή της τετραγωνικής *hinge* απώλειας, η οποία μετατρέπεται ομαλά σε μία γραμμική απώλεια στην $yf = -1$. Μπορούμε να χαρακτηρίσουμε αυτές τις συναρτήσεις απώλειας από την άποψη του τι εκτιμούν στο επίπεδο του πληθυσμού. Έστω ότι η ελαχιστοποίηση τους είναι $EL(Y, f(X))$. Στον Πίνακα 2.1 συνοψίζονται τα αποτελέσματα.

Όλες οι συναρτήσεις του παραπάνω πίνακα, εκτός από αυτή του τετραγωνικού σφάλματος, καλούνται **μεγιστοποιημένες συναρτήσεις απώλειας περιθωρίου** (Rosset et al. 2004). Αυτό σημαίνει ότι αν τα δεδομένα είναι διαχωρίσιμα, το όριο του $\hat{\beta}_\lambda$ στην (2.24), όταν $\lambda \to 0$ καθορίζει το βέλτιστο διαχωριστικό υπερεπίπεδο.

2.6 Εκτίμηση συνάρτησης και αναπαραγωγή πυρήνων

Παρακάτω γίνεται η περιγραφή των $SVMs$ μέσω της εκτίμησης παραμέτρων και της αναπαργωγής πυρήνων χώρων $Hilbert$, όπου οι ιδιότητες των πυρήνων αφθονούν.
Έστω ότι η βάση h προκύπτει από την ιδιο-επέκταση ($eigen-expansion$) ενός



Σχήμα 2.4: Η συνάρτηση απώλειας των διανυσμάτων υποστήριξης σε σύγκριση με την αρνητική απώλεια λογαριθμοπιθανοφάνειας για τη λογιστική παλινδρόμηση, με την απώλεια τετραγωνικού σφάλματος και με μία *huberized* εκδοχή της τετραγωνικής *hinge loss*.

θετικά ορισμένου πυρήνα K:

$$K(x, x') = \sum_{m=1}^{\infty} \phi_m(x)\phi_m(x')\delta_m \tag{2.25}$$

και $h_m(x) = \sqrt{\delta_m}\phi_m(x)$. Τότε για $\theta_m = \sqrt{\delta_m}\beta_m$, η (2.24) μπορεί να γραφεί ως:

$$\min_{\beta_0,\theta} \sum_{i=1}^{N} [1 - y_i(\beta_0 + \sum_{m=1}^{\infty} \theta_m \phi_m(x_i))]_+ + \frac{\lambda}{2} \sum_{m=1}^{\infty} \frac{\theta_m^2}{\delta_m}. \tag{2.26}$$

Η θεωρία της αναπαραγωγής πυρήνα χώρου *Hilbert* εγγυάται μία λύση πεπερασμένης διάστασης της μορφής:

$$f(x) = \beta_0 + \sum_{i=1}^{N} \alpha_i K(x, x_i). \tag{2.27}$$

Πίνακας 2.1: Οι *minimizers* του πληθυσμού για τις διαφορετικές συναρτήσεις απώλειας. Η λογιστική παλινδρόμηση χρησιμοποιεί τη διωνυμική λογαριθμοπιθανοφάνεια ή απόκλιση. Η γραμμική διακριτή ανάλυση χρησιμοποιεί την απώλεια του τετραγωνικού σφάλματος. Η *hinge* απώλεια του SVM εκτιμά τη λειτουργία των εκ των υστέρων πιθανοτήτων, ενώ οι άλλες εκτιμούν ένα γραμμικό μετασχηματισμό αυτών των πιθανοτήτων.

Συνάρτηση απώλειας	$L[y, f(x)]$	Συνάρτηση ελαχιστοποίησης
Διωνυμική αποκλίνουσα συμπεριφορά	$log[1 + e^{-yf(x)}]$	$f(x) = log\frac{Pr(Y=+1\|x)}{Pr(Y=-1\|x)}$
SVM Hinge απώλεια	$[1 - yf(x)]_+$	$f(x) = sign\left[Pr\left(Y=+1\|x\right) - \frac{1}{2}\right]$
Τετραγωνικό σφάλμα	$[y - f(x)]^2 = [1 - yf(x)]^2$	$f(x) = 2Pr(Y=+1\|x) - 1$
Huberized τετραγωνική *hinge* απώλεια	$-4yf(x)$, $yf(x) < -1$ $[1 - yf(x)]_+^2$, αλλιώς	$f(x) = 2Pr(Y=+1\|x) - 1$

Ο παρακάτω τύπος είναι μία ισοδύναμη μορφή του κριτηρίου βελτιστοποίησης (2.24).

$$\min_{\beta_0,\alpha} \sum_{i=1}^{N} (1 - y_i f(x_i))_+ + \frac{\lambda}{2}\alpha^T \mathbf{K}\alpha, \tag{2.28}$$

όπου το $\mathbf{K}$ είναι ένας $N \times N$ πίνακας των υπολογισμών του πυρήνα για όλα τα ζεύγη των χαρακτηριστικών και μπορεί να εκφραστεί πιο γενικά:

$$\min_{f\in\mathcal{H}} \sum_{i=1}^{N} [1 - y_i f(x_i)]_+ + \lambda J(f), \tag{2.29}$$

όπου το $\mathcal{H}$ είναι ο κατασκευαστικός χώρος των συναρτήσεων και $J(f)$ είναι ένας κατάλληλος ομαλοποιητής (*regularizer*) του χώρου. Όποιοσδήποτε από τους πυρήνες που περιγράφονται παραπάνω, μπορεί να χρησιμοποιηθεί για κάθε κυρτή συνάρτηση απώλειας και επίσης να οδηγήσει σε πεπερασμένης διάστασης αναπαράσταση του τύπου $f(x) = \beta_0 + \sum_{i=1}^{N} \alpha_i K(x, x_i)$.

2.7 Τα $SVMs$ και τα αίτια της διαστατικότητας

Έστω ότι έχουμε ένα παράδειγμα χώρου με δύο δεδομένα X_1 και X_2 και ένα πολυώνυμο πυρήνα βαθμού 2. Τότε

$$K(X, X') = (1 + \langle X, X' \rangle)^2 = (1 + X_1 X_1' + X_2 X_2')^2 \tag{2.30}$$

Αν ο αριθμός των χαρακτηριστικών p είναι πολύ μεγάλος, αλλά η κλάση διαχωρισμού προκύπτει μόνο από γραμμικό υπόχωρο διευρυμένο κατά X_1, X_2, αυτός ο πυρήνας δεν θα κατασκευαστεί εύκολα και θα έχει να ψάξει ανάμεσα σε πολλές διαστάσεις. Για να αγνοήσει κάποιος όλα τα δεδομένα εκτός από τα δύο πρώτα, θα πρέπει να γνωρίζει για τον υπόχωρο μέσα στον πυρήνα. Ο κυρίαρχος στόχος των προσαρμοστικών μεθόδων είναι να ανακαλύψουν τέτοιες κατασκευές.
Παρακάτω δίνεται ένα παράδειγμα: Έστω ότι έχουμε 2 κλάσεις με 100 παρατηρήσεις η κάθε μία. Η πρώτη κλάση έχει 4 κανονικα ανεξάρτητα χαρακτηριστικά X_1, X_2, X_3, X_4. Η δεύτερη κλάση έχει επίσης 4 κανονικα ανεξάρτητα χαρακτηριστικά για τα οποία ισχύει ότι $9 \leq \sum X_j^2 \leq 16$ και είναι επαυξημένα με 6 κανονικά *Gaussian noise* χαρακτηριστικά. Έτσι η δεύτερη ομάδα περιβάλλει την πρώτη όπως η φλούδα το πορτοκάλι σε ένα 4-διάστατο υπόχωρο. Δίνονται παρατηρήσεις από 1000 παρατηρήσεις ελέγχου και συγκρίνουμε τις διαφορετικές μεθόδους. Στον Πίνακα 2.2 φαίνονται τα αποτελέσματα:

Πίνακας 2.2: Στον πίνακα παρουσιάζονται οι μέσες τιμές των σφαλμάτων ελέγχου και σε παρένθεση τα τυπικά σφάλματα των μέσων για 50 προσομοιώσεις.

		Test Error (SE)	
	Method	No Noise Features	Six Noise Features
1	SV Classifier	0.450 (0.003)	0.472 (0.003)
2	SVM/poly 2	0.078 (0.003)	0.152 (0.004)
3	SVM/poly 5	0.180 (0.004)	0.370 (0.004)
4	SVM/poly 10	0.230 (0.003)	0.434 (0.002)
5	BRUTO	0.084 (0.003)	0.090 (0.003)
6	MARS	0.156 (0.004)	0.173 (0.005)
	Bayes	0.029	0.029

Στην πρώτη γραμμή χρησιμοποιείται ο ταξινομητής διανυσμάτων υποστήριξης στο αρχικό χώρο χαρακτηριστικών. Οι γραμμές 2-4 αναφέρονται στις μηχανές διανυσμάτων υποστήριξης με 2-,5- και 10-διάστατο πυρήνα πολυώνυμο. Για όλες τις μεθόδους διανυσμάτων υποστήριξης, η παράμετρος κόστους C είναι

επιλεγμένη έτσι ώστε να ελαχιστοποιεί το σφάλμα. Η γραμμή 5 προσαρμόζει ένα πρόσθετο *spline* μοντέλο στην (-1,+1) απόκριση των ελάχιστων τετραγώνων, χρησιμοποιώντας τον αλγόριθμο του $BRUTO$ για πρόσθετα μοντέλα, που περιγράφηκε από το βιβλίο του Hastie and Tibshirani (1990). Η γραμμή 6 χρησιμοποιεί τα $MARS$, (*multivariate adaptive regression splines*), το οποίο είναι συγκρίσιμο με το SVM πολυώνυμο βαθμού 10. Η μέθοδος του $BRUTO$ και $MARS$ έχουν την ικανότητα να αγνοούν τις περιττές μεταβλητές.

Στον αρχικό χαρακτηριστικό χώρο ένα υπερεπίπεδο δε μπορεί να διαχωρίσει τις κλάσεις και ο SV ταξινομητής το κάνει με πολλά σφάλματα. Οι πολυωνυμικοί SVM έχει μία ουσιαστική βελτίωση στο σφάλμα αλλά επηρεάζεται από τα 6 χαρακτηριστικά θορύβου *noise*. Είναι επίσης πολύ ευαίσθητο στην επιλογή του πυρήνα, με την επιλογή του πυρήνα βαθμού 2 να έχει τα καλύτερα αποτελέσματα.

2.8 Οι μηχανές διανυσμάτων υποστήριξης στην παλινδρόμηση

Έστω ότι έχουμε το γραμμικό μοντέλο παλινδρόμησης:

$$f(x) = x^T\beta + \beta_0 \tag{2.31}$$

Για την εκτίμηση του β, ελαχιστοποιούμε το:

$$H(\beta, \beta_0) = \sum_{i=1}^{N} V(y_i - f(x_i)) + \frac{\lambda}{2}\|\beta\|^2, \tag{2.32}$$

όπου

$$V_\epsilon(r) = \begin{cases} 0, & \text{αν } \mid r \mid < \epsilon, \\ \mid r \mid -\epsilon, & \text{αλλιώς.} \end{cases}$$

Αυτός είναι ένας μετρητής σφάλματος που εξαρτάται από το ϵ. Υπάρχει μία αναλογία με τη διάταξη του διαχωριστή διανυσμάτων υποστήριξης, όπου τα σημεία που βρίσκονται στη σωστή πλευρά του ορίου και μακριά από αυτό αγνοούνται κατά τη βελτιστοποίηση. Στην παλινδρόμηση, αυτά τα σημεία μικρού σφάλματος είναι αυτά με τα μικρά υπόλοιπα.

Αν τα $\hat{\beta}, \hat{\beta}_0$ ελαχιστοποιούν την H, η λύση της συνάρτησης παίρνει τη μορφή:

$$\hat{\beta} = \sum_{i=1}^{N} (\hat{\alpha}_i^* - \hat{\alpha}_i) x_i \tag{2.33}$$

$$\hat{f}(x) = \sum_{i=1}^{N} (\hat{\alpha}_i^* - \hat{\alpha}_i)\langle x, x_i\rangle + \beta_0, \tag{2.34}$$

όπου τα $\hat{\alpha}_i^*, \hat{\alpha}_i$ είναι θετικά και λύνουν το παρακάτω πρόβλημα τετραγωνικού προγραμματισμού.

$$\min_{\alpha_i^*, \alpha_i} \epsilon \sum_{i=1}^{N} (\alpha_i^* + \alpha_i) - \sum_{i=1}^{N} y_i(\alpha_i^* - \alpha_i) + \frac{1}{2} \sum_{i,i'=1}^{N} (\alpha_i^* - \alpha_i)(\alpha_{i'}^* - \alpha_{i'}) \langle x_i, x_{i'} \rangle \quad (2.35)$$

δεδομένων των περιορισμών

$$0 \leq \alpha_i, \alpha_i^* \leq 1/\lambda$$
$$\sum_{i=1}^{N} y_i(\alpha_i^* - \alpha_i) = 0$$
$$\alpha_i \alpha_i^* = 0.$$

Λόγω της φύσης αυτών των περιορισμών, τυπικά μόνο ένα υποσύνολο των τιμών λύσης $(\hat{\alpha}_i^* - \hat{\alpha}_i)$ είναι μη μηδενικό, και οι σχετικές τιμές καλούνται διανύσματα υποστήριξης. Η λύση εξαρτάται από τα δεδομένα εισαγωγής μόνο μέσω του εσωτερικού γινομένου $\langle\ x_i, x_{i'}\ \rangle$. Έτσι μπορούμε να γενικεύσουμε τη μέθοδο σε μικρότερους χώρους, καθορίζοντας ένα κατάλληλο εσωτερικό γινόμενο.

Σημειώνεται ότι οι παράμετροι ϵ και λ σχετίζονται με το κριτήριο (2.32). Το ϵ είναι η παράμετρος της συνάρτησης απώλειας V_ϵ και η παράμετρος λ είναι μία πιο παραδοσιακή παράμετρος κανονικοποίησης.

3 MSPC, SVM και αναγνώριση μοτίβων

3.1 Εισαγωγή

Μετά από μία εισαγωγή στις βασικές αρχές του μονομεταβλητού και του πολυμεταβλητού στατιστικού ελέγχου διεργασιών και στις $SVMs$, μπορεί κανείς να κατανοήσει τη σπουδαιότητα των διαγραμμάτων ελέγχου του ΣΕΔ στην παρακολούθηση των ποιοτικών χαρακτηριστικών μιας κατασκευαστικής διεργασίας. Η εμφάνιση ενός μη φυσικού μοτίβου σε ένα διάγραμμα ελέγχου οδηγεί στο συμπέρασμα ότι η διεργασία έχει επηρεαστεί από προσδιορισμένα αίτια, και πρέπει να γίνουν διορθωτικές ενέργειες. Η αναγνώριση των διαφορετικών και ανεξάρτητων μοτίβων είναι ένα περίπλοκο και μη γραμμικό πρόβλημα ταξινόμησης. Στις μονομεταβλητές διεργασίες το $Shewhart$ διάγραμμα είναι κατάλληλο για να μας δείξει τα ειδικά αίτια μεταβλητότητας στη διεργασία αλλά στις περισσότερες περιπτώσεις οι διεργασίες αποτελούνται από παραπάνω μεταβλητές. Η αναγνώριση μοτίβων σε διαγράμματα ελέγχου αναφέρεται στη χρήση ορισμένων μεθόδων ανάλυσης που κατεργάζονται και αναλύουν τα ποιοτικά χαρακτηριστικά μιας χρονικής σειράς δεδομένων, συμπεριλαμβάνονται πληροφορίες για την κατάσταση των χαρακτηριστικών και την τάση μεταβολών στη διαδικασία.

Υπάρχουν πολλές έρευνες με τα αποτελέσματά τους στη βιβλιογραφία, σχετικά με την αναγνώριση μη φυσικών μοτίβων στα πολυμεταβλητά διαγράμματα ελέγχου. Τα παλαιότερα χρόνια η αναγνώριση τέτοιων μοτίβων εξαρτιόταν από τις ικανότητες και την εμπειρία που είχε ο αναλυτής. Αρκετές μελέτες ασχολήθηκαν με συμπληρωματικούς κανόνες στην ερμηνεία των διαγραμμάτων ελέγχου και στην αναζήτηση προσδιορισμένων αιτιών. Το κύριο μειονέκτημα αυτών των ελέγχων είναι ότι αναγνωρίζουν την ύπαρξη ενός μη φυσικού μοτίβου, αλλά δεν δείχνουν ρητά ποιο είναι. Επίσης η χρήση πρόσθετων ελέγχων αυξάνει το ρίσκο λανθασμένου συναγερμού.

Σε πολλές βιομηχανικές διαδικασίες, υπάρχουν πολλές μεταβλητές που ορίζουν την γενική ποιότητα (*overall quality*) ενός προϊόντος, η οποία συνήθως καθορίζεται από το επίπεδο των ποιοτικών χαρακτηριστικών. Αυτά τα ποιοτικά χαρακτηριστικά μπορεί να είναι συσχετισμένα και τα ξεχωριστά διαγράμματα ελέγχου για τη παρακολούθηση των μεμονωμένων ποιοτικών χαρακτηριστικών να μην επαρκούν για την ανίχνευση αλλαγών στη γενική ποιότητα του προϊόντος. Έτσι είναι σημαντική η ύπαρξη διαγραμμάτων ελέγχου που μπορούν να παρακολουθήσουν ταυτόχρονα πολυμεταβλητές μετρήσεις. Διάφορα πρόσφατα άρθρα προτείνουν την εφαρμογή διαδικασιών ελέγχου (π.χ το $Hotelling$ T^2 διάγραμμα, το $MCUSUM$ διάγραμμα και το $MEWMA$ διάγραμμα) για την παρουσίαση ενός εκτός ελέγχου σήματος σε ένα πολυμεταβλητό διάγραμμα ελέγχου. Στα πολυμεταβλητά διαγράμματα ελέγχου μπορεί να εμφανιστούν

πολλά διαφορετικά συστηματικά μοτίβα και το κύριο μειονέκτημα των περισσότερων πολυμεταβλητών διαδικασιών των διαγραμμάτων ελέγχου είναι ότι δεν μπορούν κατευθείαν να βρουν πιο μη φυσικό μοτίβο συμβαίνει.

Αν συγκρίνουμε την απόδοση των $MSVMs$ ($Multiclass\ Support\ Vector\ Machines$) στην ακρίβεια της ταξινόμησης με αυτή κάποιων παλαιότερων δικτύων, όπως το δίκτυο της προς τα πίσω διάδοσης ($back-propagation\ network$- BPN) ή τα δίκτυα εκμάθησης κβαντικού διανύσματος ($learning\ vector\ quantization$- LVQ) και $LVQ-x$, μέσω αριθμητικής προσομοίωσης, τα $MSVMs$ εμφανίζουν πολύ καλύτερα αποτελέσματα και σε αυτά θα γίνει αναφορά παρακάτω.

3.2 SVM για αναγνωριση μοτίβων διαγραμμάτων ελέγχου σε πολυμεταβλητές διεργασίες

Στη συνέχεια προτείνεται μία προσέγγιση με SVM για την αναγνώριση μη φυσικών μοτίβων σε πολυμεταβλητές διεργασίες. Να σημειωθεί ότι το SVM έχει παρουσιαστεί σαν μία νέα τεχνική για τη λύση μιας ποικιλίας μαθησιακών προβλημάτων, προβλημάτων ταξινόμησης και πρόβλεψης.

Εκτελούνται δύο ταξινομητές που βασίζονται στο SVM και στην διακριτική ανάλυση ($discriminant\ analysis$) για να αναγνωρίσουν πολυμεταβλητά μη φυσικά μοτίβα. Η διακριτική ανάλυση είναι μία πολυμεταβλητή τεχνική που ασχολείται με ευκρινώς διαχωριζόμενες ομάδες παρατηρήσεων και καινούριες παρατηρήσεις κατανεμημένες σε προηγούμενες καθορισμένες ομάδες και χρησιμοποιείται σαν βάση για τη σύγκριση. Επιπλέον προτείνονται και δύο εναλλακτικές διεργασίες ταξινόμησης για ταξινομητή που βασίζεται στο SVM. Παρακάτω γίνεται μία παρουσίαση του ταξινομητή βάσει των SVM.

Οι μηχανές διανυσμάτων υποστήριξης παρουσιάζουν αξιοσημείωτες ιδιότητες και ικανότητες γενίκευσης σε περιπτώσεις ταξινόμησης και ανάλυσης. Είναι ένα δυνατό εργαλείο μάθησης μηχανών που είναι ικανό να παρουσιάζει μη γραμμικές σχέσεις και να παράγει μοντέλα που γενικεύουν τα αφανή δεδομένα. Στην πράξη η ταξινόμηση συνήθως περιλαμβάνει δεδομένα εκπαίδευσης και ελέγχου ($test$), τα οποία περιλαμβάνουν κάποια δείγματα δεδομένων ($data\ instances$). Κάθε δείγμα σε μία ομάδα εκπαίδευσης περιέχει μία τιμή-στόχο (ετικέτα κλάσης) και διάφορα γνωρίσματα ($attributes$). Στόχος είναι ο ταξινομητής να παράγει ένα μοντέλο που προβλέπει την τιμή-στόχο των δειγμάτων σε ένα σετ που δίνονται μόνο τα γνωρίσματα. Το βασικό σκεπτικό του SVM είναι να μεταφέρει τα δεδομένα σε ένα χώρο μεγαλύτερης διάστασης και να βρει το βέλτιστο υπερεπίπεδο στο χώρο που μεγιστοποιεί το περιθώριο μεταξύ των κλάσεων. Το κύριο αντικείμενο της εφαρμογής του SVM στη λύση προβλημάτων ταξινόμησης περιλαμβάνει δύο στάδια. Στο πρώτο, το SVM μετατρέπει το χώρο εισαγωγής

σε ένα χώρο μεγαλύτερης διάστασης μέσω μίας μη-γραμμικής συνάρτησης αντιστοίχισης (*mapping*) και, στο δεύτερο στάδιο κατασκευάζει ένα διαχωριστικό υπερεπίπεδο με μέγιστη απόσταση από τα κοντινότερα σημεία των δεδομένων εκπαίδευσης. Στο SVM ο τύπος της συνάρτησης πυρήνα και άλλες παράμετροι έχουν μεγάλη επίδραση στην απόδοση.

3.3 Παραγωγή ενός συνόλου δεδομένων εκπαίδευσης σε πολυμεταβλητές διεργασίες

Σε πολλές βιομηχανίες υπάρχουν κάποια ποιοτικά χαρακτηριστικά που πρέπει να παρακολουθούνται ταυτόχρονα. Το στατιστικό που χρησιμοποιείται συνήθως είναι το

$$\chi_t^2 = n(\bar{\mathbf{X}}_t - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}}_t - \boldsymbol{\mu}_0), \tag{3.1}$$

όπου τα $\bar{\mathbf{X}}_t$ αναπαριστούν το μέσο των ποιοτικών χαρακτηριστικών που παρακολουθούνται, $\boldsymbol{\mu}_0 = \mathbf{0}$ και $\boldsymbol{\Sigma}$ ένας γνωστός πίνακας διασποράς. Όταν το διάνυσμα των μέσων και ο πίνακας διασποράς είναι άγνωστα, τα αντικαθιστούμε με τους εκτιμητές $\bar{\bar{\mathbf{X}}}$ και $\mathbf{S}$, αντίστοιχα. Το στατιστικό ελέγχου για κάθε δείγμα μπορεί να γραφεί ως

$$T_t^2 = n(\bar{\mathbf{X}}_t - \bar{\bar{\mathbf{X}}})'\mathbf{S}^{-1}(\bar{\mathbf{X}}_t - \bar{\bar{\mathbf{X}}}). \tag{3.2}$$

Όταν χρησιμοποείται αυτός ο τύπος η διαδικασία ελέγχου καλείται συνήθως *Hotelling* T^2 διάγραμμα ελέγχου (όπως αναφέρεται και στο 1ο κεφάλαιο). Πρακτικά δεν είναι διαθέσιμα αρκετά παραδείγματα εκπαίδευσης από μη φυσικά μοτίβα. Μία συνήθης προσέγγιση που ακολουθούν άλλοι ερευνητές είναι η παραγωγή παραδειγμάτων εκπαίδευσης βασισμένοι σε ένα προκαθορισμένο μαθηματικό μοντέλο (Cheng (1997), Guh and Hsieh (1999)). Μία p-διάστατη πολυμεταβλητή κανονική διεργασία προσομοιώνεται με την παραγωγή ψευδοτυχαίων μεταβλητών από μία πολυμεταβλητή κανονική κατανομή με μέσο μ και πίνακα διασποράς έναν $p \times p$ πίνακα (για λόγους ευκολίας ο $\boldsymbol{\Sigma}$ είναι γνωστός). Ένα φυσικό ή τυχαίο μοτίβο θα παραχθεί από τον γενικό τύπο:

$$\mathbf{X}_t = \boldsymbol{\mu} + \mathbf{R}_t, \tag{3.3}$$

όπου $\mathbf{X}_t$ είναι τα p ποιοτικά χαρακτηριστικά τη χρονική στιγμή t, $\boldsymbol{\mu}$ είναι ο γνωστός μέσος των σειρών δεδομένων όταν η διεργασία είναι εντός ελέγχου και $\mathbf{R}_t$ είναι ο τυχαίος θόρυβος τη χρονική στιγμή t. Όταν εμφανιστεί ένα μοτίβο, εισάγεται μία νέα συνιστώσα θορύβου d_t στο μέσο της διεργασίας από τις εσφαλμένες μεταβλητές. Η τιμή d_t υποδηλώνει μία διατάραξη τη στιγμή t λόγω μη προσδιορισμένων αιτιών. Με τη χρήση αυτής της μεταβλητής μπορεί

να προσομοιωθεί ένα μοτίβο. Στην έρευνα (Cheng and Cheng, (2007)) που περιγράφεται παρακάτω εξετάζονται 4 συνήθη μη φυσικά μοτίβα στα διαγράμματα ελέγχου, τάση ($trend$), ξαφνικές αλλαγές, μίξεις και κυκλικά μοτίβα.

1. Οι τάσεις μπορούν να εκφραστούν ως d_t=θt, όπου θ είναι η τάση της κλίσης συναρτήσει του σ.

2. Οι ξαφνικές αλλαγές μπορούν να γραφτούν ως d_t=$u\delta$, όπου το u αναπαριστά την παράμετρο που καθορίζει την εμφάνιση της αλλαγής, $u = 0$ όταν δεν υπάρχουν αλλαγές και $u = 1$ στην περίπτωση αλλαγής, αντίστοιχα. Η ποσότητα δ μπορεί να οριστεί ως το μέγεθος της αλλαγής του μέσου συναρτήσει του σ.

3. Οι μίξεις μπορούν να εκφραστούν ως d_t=$(-1)^{\varepsilon}\Delta$, όπου ε είναι η παράμετρος που καθορίζει την εμφάνιση αλλαγής μεταξύ των κατανομών. Η ποσότητα μπορεί να ελεγχθεί από έναν τυχαίο αριθμό v (με τιμές μεταξύ του 0 και του 1) και από την πιθανότητα της αλλαγής μεταξύ των κατανομών (Pr). Η ποσότητα $\varepsilon = 0$ αν $v < Pr$ και $\varepsilon = 1$ αν $v \geq Pr$. Παρακάτω θεωρούμε $Pr = 0.5$. Η ποσότητα Δ μπορεί να οριστεί ως το αντιστάθμισμα από το μέσο της διαδικασίας συναρτήσει του σ.

4. Τα κυκλικα μοτίβα μπορούν να μοντελοποιηθούν ως d_t=$\kappa sin(2\pi t/\Omega)$, όπου κ είναι το εύρος των κυκλικών μοτίβων συναρτήσει του σ, και το σύμβολο Ω καθορίζεται από την περίοδο του κυκλικού μοτίβου (εδώ $\Omega =$ 16).

Η προσομοίωση εφαρμόζεται με τη χρήση του λογισμικού $Matlab$ και για απλότητα οι μεταβλητές έχουν μέσο 0 και διασπορά 1. Έτσι ο πίνακας διασποράς έχει τα στοιχεία της διαμέσου του ίσα με 1 και τα υπόλοιπα είναι σε ζευγάρια. Στον Πίνακα 3.1 συνοψίζονται οι παράμετροι που χρησιμοποιούνται για την προσομοίωση των μοτίβων. Τα δεδομένα που είναι εντός ελέγχου περιλαμβάνονται στην ομάδα εκπαίδευσης. Εξετάζονται επίσης πολυμεταβλητές διεργασίες με διάφορους θετικούς και αρνητικούς συντελεστές συσχέτισης για να καλύψουν το εύρος των παραμέτρων για κάθε τύπο μοτίβου.

Σε αυτή την έρευνα Cheng and Cheng (2007), οι ερμηνείες των λανθασμένων τιμών περιγράφονται όπως παρακάτω. Έστω p=3, η εμφάνιση μη φυσικών μοτίβων προκύπτει με έναν από τους παρακάτω τρόπους:

1. Μόνο μία λάθος μεταβλητή. Μόνο μία από τις μεταβλητές προκαλεί το μη φυσικό μοτίβο.

2. Δύο λάθος μεταβλητές. Δύο από τις μεταβλητές προκαλούν μη φυσικό μοτίβο, όσο οι υπόλοιπες μεταβλητές παραμένουν σε κατάσταση εντός ελέγχου.

Πίνακας 3.1: Οι παράμετροι των παραδειγμάτων εκπαίδευσης για μη φυσικά μοτίβα

Pattern type	Parameters	No. of training cases
Natural	In-control data	3000
Trend	gradient: 0.10, 0.125, 0.15	3000
Sudden shift	shift magnitude: 1.5, 2.0, 3.0	3000
Mixture	magnitude: 2.0, 2.5, 3.0	3000
Cyclic pattern	amplitude: 1.5, 2.0, 3.0; period: 16	3000

3. Και οι τρεις μεταβλητές λάθος. Όλες οι μεταβλητές προκαλούν τον ίδιο τύπο μη φυσικού μοτίβου.

- Επιλογή εισαχθέντος διανύσματος:
  Η αναπαράσταση των δεδομένων στο σύνολο εκπαίδευσης έχει έντονη επιρροή στην απόδοση του ταξινομητή. Ο προσδιορισμός ενός επαρκούς μεγέθους παραθύρου είναι ένα από τα πιο σημαντικά βήματα στην παρούσα εφαρμογή (Cheng and Cheng (2007)). Αφού οι γρήγοροι υπολογισμοί είναι το πιο σημαντικό κομμάτι για τον έλεγχο των διεργασιών, είναι καθοριστικό να ελαττωθεί το μέγεθος του παραθύρου που εξασφαλίζει αποτελεσματικούς υπολογισμούς. Προηγούμενες μελέτες δείχνουν ότι ένα μικρό μέγεθος παραθύρου μπορεί να προκαλέσει ένα μεγαλύτερο σφάλμα τύπου I, εξαιτίας των ανεπαρκών πληροφοριών των χαρακτηριστικών των δεδομένων. Από την άλλη το μεγάλο μέγεθος του παραθύρου απαιτεί μεγάλο χρόνο υπολογισμών. Στην παρούσα έρευνα το μέγεθος του παραθύρου έχει επιλεχθεί μέσω πειραμάτων να είναι 32, έτσι χρειάζεται να υπολογιστούν 32 T^2 στατιστικά όσα και τα συστατικά του χαρακτηριστικού διανύσματος.

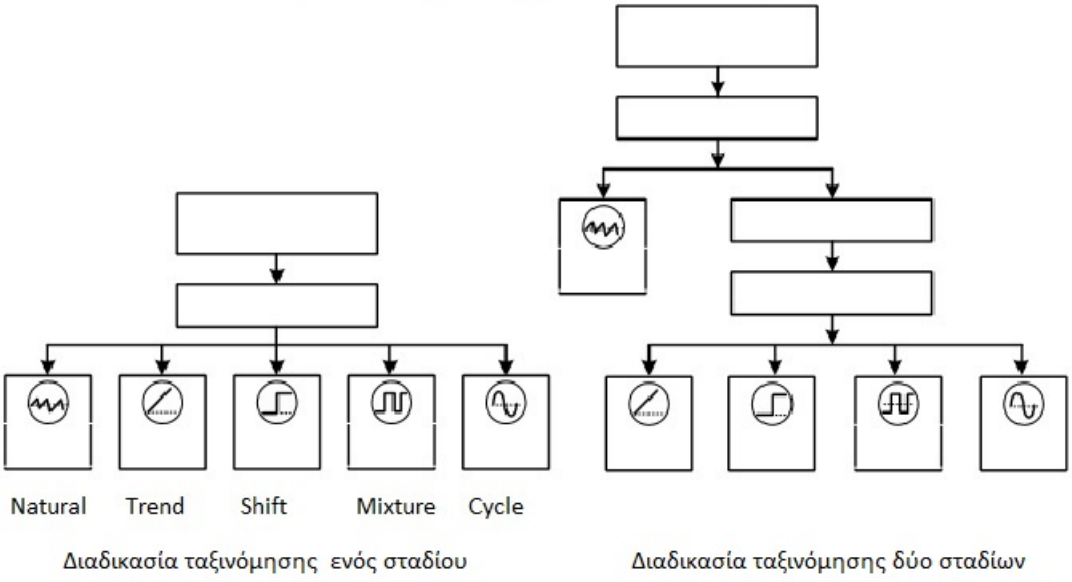
- Επιλογή μοντέλου και εκπαίδευση:
  Αν και οι μηχανές διανυσμάτων υποστήριξης έχουν καλή επίδοση σε έναν αριθμό εφαρμογών, ένα πρόβλημα που θα αντιμέτωπίσει ο χρήστης ενός SVM είναι η επιλογή του πυρήνα και ειδικά των παραμέτρων για αυτόν τον πυρήνα. Οι συναρτήσεις πυρήνα αντιστοιχίζουν τα αρχικά δεδομένα σε ένα χώρο μεγαλύτερης διάστασης όπου η ομάδα των δεδομένων που εισάγονται διαχωρίζεται γραμμικά. Η επιλογή των συναρτήσεων πυρήνα είναι το σημαντικότερο πρόβλημα και ο πιο σημαντικός παράγοντας στις εφαρμογές των μηχανών διανυσμάτων υποστήριξης. Στο συγκεκριμένο πρόβλημα ταξινόμησης γίνεται σύγκριση μεταξύ των 4 διαφορετικών συναρτήσεων πυρήνα από τον Πίνακα 3.1 βάσει της απόδοσης τους στην τα-

ξινόμηση για την επιλογή της συνάρτησης πυρήνα. Συνεπώς η καλύτερη απόδοση εξασφαλίζεται με τη συνάρτηση ακτινικής βάσης και είναι η πιο δημοφιλής επιλογή τύπων πυρήνα κάνοντας χρήση των SVM, και αυτή θα χρησιμοποιηθεί παρακάτω. Για αυτή τη συνάρτηση πυρήνα πρέπει να καθοριστούν η παράμετρος πυρήνα γ και ο συντελεστής ποινής C. Οι παράμετροι πυρήνα καθορίζουν την κατασκευή του χαρακτηριστικού χώρου υψηλής διάστασης, όπου θα βρεθεί ένα μέγιστο περιθώριο υπερεπιπέδου. Αυτές οι παράμετροι επιλέγονται συνήθως μέσω πειράματος. Όσο μεγαλύτερη είναι η παράμετρος C, τόσο μεγαλύτερη είναι η ποινή των λαθών και πρέπει να επιλεχθεί με προσοχή έτσι ώστε να αποφευχθεί η υπερπροσαρμογή ($overfitting$). Στο συγκεκριμένο πείραμα οι παράμετροι γ και C βρίσκονται μεταξύ των ευρών [0.03125, 2] και [2, 40] αντίστοιχα.

- Δύο διαδικασίες για την ανακάλυψη σημάτων εκτός ελέγχου και την ταξινόμηση:
  Οι (Cheng and Cheng (2007)) πρότειναν δύο εναλλακτικά σχέδια για την κατασκευή αναγνωριστών πολυμεταβλητών μη φυσικών μοτίβων. Το πρώτο σχέδιο, το οποίο αναφέρεται και ως διαδικασία ταξινόμησης ενός σταδίου ($one\ stage\ classification\ procedure$), περιλαμβάνει ένα ταξινομητή που βασίζεται στο SVM, ο οποίος είναι σχεδιασμένος να αναγνωρίζει μοτίβα σε πολυμεταβλητές διεργασίες σε ένα συγκεκριμένο χρόνο. Τα δεδομένα που εισάγονται περιλαμβάνουν εντός και εκτός ελέγχου δεδομένα.
  Το δεύτερο σχέδιο, το οποίο αναφέρεται και ως διαδικασία ταξινόμησης δύο σταδίων ($two\ stage\ classification\ procedure$), περιέχει δύο $SVMs$ σε σειρά (τις SVM-Ι και SVM-ΙΙ). Το $SVM-I$ λειτουργεί σαν ανιχνευτής. Όταν ανιχνεύεται ένα εκτός ελέγχου σήμα, ο $SVM-II$ ταξινομητής λειτουργεί ώστε να ανακαλύψει τον τύπο του μη φυσικού μοτίβου για το σήμα που ανιχνεύθηκε. Στο Σχήμα 3.1 φαίνεται η αρχιτεκτονική των δύο διαδικασιών.

3.4 Αποτελέσματα

Κατά τη διαδικασία της αξιολόγησης, υπολογίζεται ο βαθμός της σωστής ταξινόμησης $ROCC$ ($Rate\ Of\ Correct\ Classification$). Τα δείγματα ελέγχου παράγονται με τον ίδιο τρόπο. Κάθε ένας από τους ταξινομητές ελέγχεται με 75000 δείγματα ελέγχου. Παρόμοια κάθε τύπος μοτίβου ελέγχεται με διάφορους θετικούς και αρνητικούς συντελεστές συσχέτισης σε πολυμεταβλητές



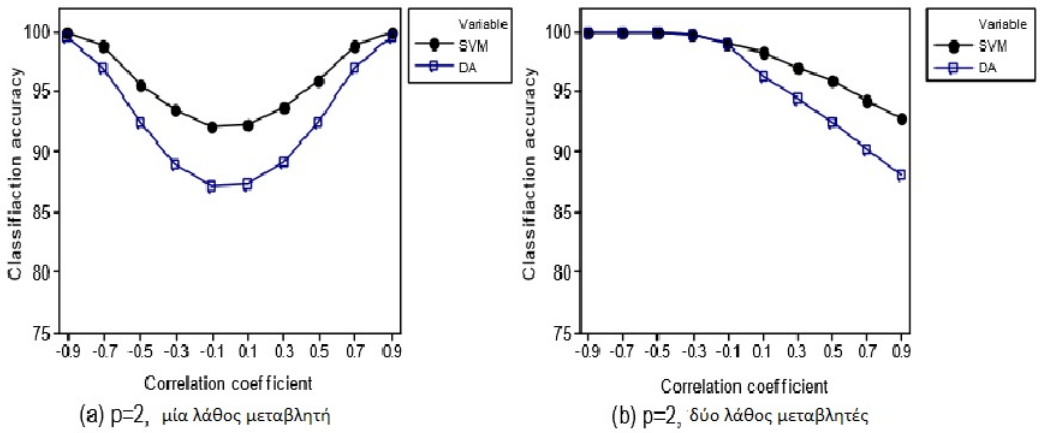
Σχήμα 3.1: Η αρχιτεκτονική των δύο διαδικασιών.

διεργασίες.

Η απόδοση των διαφορετικών ταξινομητών:

Οι δύο ταξινομητές μεταξύ των οποίων γίνεται η σύγκριση βασίζονται στο SVM και στην διακριτική ανάλυση και εισάγονται τα ίδια δεδομένα. Σαν βάση επιλέγεται η απόδοση της διακριτικής ανάλυσης και τα αποτελέσματα αναφέρονται για την διαδικασία ταξινόμησης ενός σταδίου με μεταβλητή ρ. Στο Σχήμα 3.2 εμφανίζεται η γενική απόδοση του SVM και της διακριτικής ανάλυσης για τη διαδικασία ταξινόμησης ενός σταδίου.

Παρατηρούμε ότι η απόδοση των ταξινομητών που βασίζονται στο SVM έχουν πιο σταθερά αποτελέσματα από αυτούς που βασίζονται στη διακριτική ανάλυση με θετικούς και αρνητικούς συντελεστές συσχέτισης. Αξίζει να σημειωθεί ότι οι μέθοδοι των SVM και της διακριτικής ανάλυσης παρέχουν καλύτερη απόδοση όταν ρ=-0.9 και ρ=0.9 και ότι η χαμηλότερη απόδοση παρατηρείται για ρ=-0.1 και ρ=0.1 . Επίσης από το διάγραμμα φαίνεται ότι ο αναγνωριστής μοτίβου που βασίζεται στο SVM έχει την ίδια ικανότητα στην ανίχνευση διαφορετικών διευθύνσεων αλλαγών του ρ. Στην περίπτωση των δύο λανθασμένων μεταβλητών η απόδοση του ταξινομητή βελτιώνεται όσο η τιμή του ρ μειώνεται. Οπότε μπορούμε να συμπεράνουμε ότι οι ταξινομητές που βασίζονται στα SVM υπερτερούν σε σχέση με τις τεχνικές της παραδοσιακής διακριτικής ανάλυσης.



Σχήμα 3.2: Γενική απόδοση για τη διαδικασία ταξινόμησης ενός σταδίου.

Σύγκριση διαφορετικών διαδικασιών ταξινόμησης:

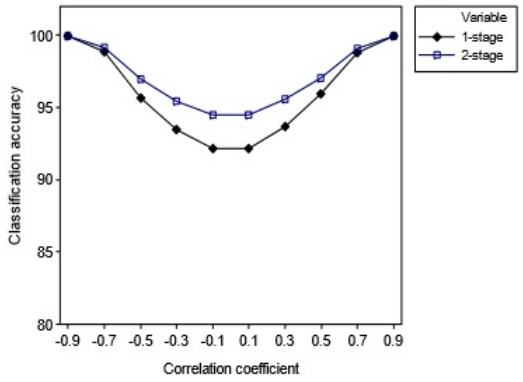
Στο Σχήμα 3.3 παρουσιάζεται η γενική απόδοση των SVM για δύο τύπους διαδικασίας ταξινόμησης σε διδιάστατη διεργασία (p=2). Για την περίπτωση που έχουμε μία λανθασμένη τιμή, παρατηρούμε ότι η διαδικασία ταξινόμησης σε δύο στάδια έχει καλύτερα αποτελέσματα από την προσέγγιση με ένα στάδιο.

Όπως δείχνει και το Σχήμα 3.3 ο αναγνωριστής μοτίβων με βάση το SVM έχει την ίδια ικανότητα να ανιχνεύει αλλαγές του ρ σε διαφορετικές κατευθύνσεις. Για p=3 τα αποτελέσματα δίνονται στον Πίνακα 3.2. Είναι εμφανές ότι η διαδικασία ταξινόμησης σε δύο στάδια συμπεριφέρεται σημαντικά καλύτερα από την προσέγγιση σε ένα στάδιο.

Πίνακας 3.2: Γενική απόδοση των $SVMs$ για διαφορετικές διαδικασίες ταξινόμησης, p=3.

Unnatural type	ρ	One-stage classification procedure	Two-stage classification procedure	Increment %
One errant variable only	0.1	90.39 / 90.34	93.27 / 93.06	2.88 / 2.72
	0.3	92.87 / 92.55	95.07 / 94.79	2.20 / 2.24
	0.5	96.04 / 95.86	97.39 / 97.15	1.35 / 1.29
	0.7	99.12 / 99.10	99.37 / 99.31	0.25 / 0.21
	0.9	100.00 /100.00	100.00 /100.00	0.00 / 0.00
All three errant variables	0.1	99.13 / 99.15	99.30 / 99.36	0.17 / 0.21
	0.3	97.98 / 97.56	98.49 / 98.20	0.51 / 0.64
	0.5	95.63 / 95.56	96.93 / 96.73	1.30 / 1.17
	0.7	93.26 / 92.98	95.07 / 94.94	1.81 / 1.96
	0.9	90.53 / 90.28	92.93 / 92.89	2.40 / 2.61

Note: (a/b) denotes (training/test).

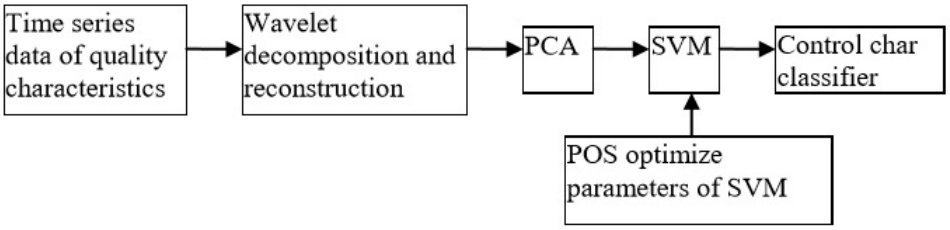


Σχήμα 3.3: Γενική απόδοση των $SVMs$ για διαφορετικές διαδικασίες ταξινόμησης, p=2

3.5 Αναγνώριση υβριδικών μοτίβων που βασίζεται σε $WA-PCA-PSO-SVM$

Ακόμα καλύτερα αποτελέσματα στην αναγνώριση μοτίβων στα διαγράμματα ελέγχου έχουμε αν σε ένα μοντέλο συνδυάσουμε την κυματική ανάλυση ($Wavelet\ Analysis-WA$), την ανάλυση κύριων συνιστωσών ($Principal\ Component\ Analysis-PCA$), τη βελτιστοποίηση πλήθους σωματιδίων ($Particle\ Swarm\ Optimization-PSO$) και τις SVM. Η WA είναι κατάλληλη στο να εξαλείψει τα δεδομένα που προκαλούν θορύβους στο διάγραμμα ελέγχου και έχουν δυσμενείς επιπτώσεις στην αναγνώριση μοτίβων. Η PCA εξαλείφει τις περιττές πληροφορίες των δεδομένων μεταξύ των SVM και μειώνει τη διάσταση των εισαχθέντων δεδομένων και την υπολογιστική πολυπλοκότητα. Ο αλγόριθμος βελτιστοποίησης πλήθους σωματιδίων βελτιώνει τις παραμέτρους του SVM, η σταθεροποίηση του βέλτιστου ταξινομητή μοτίβων διαγραμμάτων ελέγχου μπορεί να λύσει το πρόβλημα των βέλτιστων παραμέτρων του SVM και τέλος η συλλογή των σημείων χωρίζεται απο το SVM. Τα αποτελέσματα προσομοιώσεων που έχουν γίνει (Yan-Zhong et al. (2014), δείχνουν ότι αυτό το μοντέλο είναι εφικτό και τα αποτελέσματά του είναι αξιόπιστα. Βελτιώνει την ακρίβεια στην αναγνώριση μη φυσικών μοτίβων και χρησιμοποιείται στην παρακολούθηση μηχανικών διαδικασιών σε πραγματικό χρόνο. Στο Σχήμα 3.4 φαίνεται ο τρόπος αναγνώρισης μοτίβων ενός διαγράμματος ελέγχου με αυτόν τον αλγόριθμο.

Το SVM μπορεί να λύσει δύο προβλήματα ταξινόμησης. Η αναγνώριση διαγραμμάτων ελέγχου είναι ένα είδος πολυ-ταξινόμησης προβλημάτων, επομένως ο ταξινομητής πρέπει να κατασκευαστεί με τη στρατηγική του συνδυασμού.



Σχήμα 3.4: Μοντέλο αναγνώρισης διαγραμμάτων ελέγχου

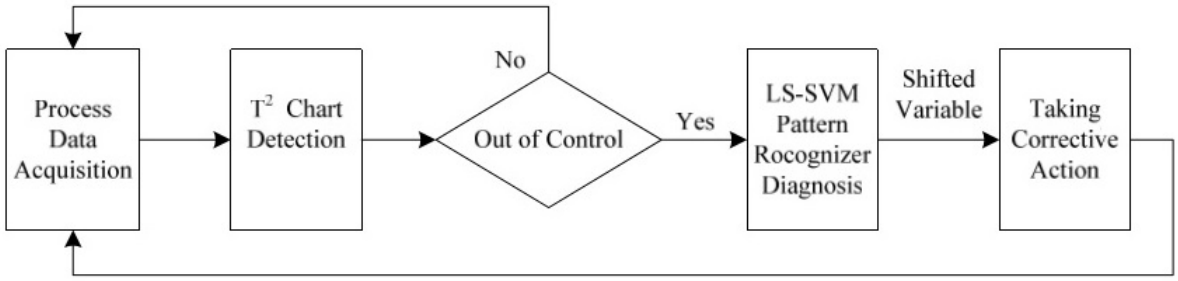
Υιοθετούμε το κατευθυνόμενο ακυκλικό γράφημα (*Directed Acyclic Graph-DAG*) που συνδυάζει δύο ομάδες ταξινομητών και λύνει το πρόβλημα ταξινόμησης πολλών κλάσεων. Το χρησιμοποιούμε για να βελτιώσουμε τον αλγόριθμο PSO που συνδυάζει τα χαρακτηριστικά διανύσματα και τις παραμέτρους της συνάρτησης πυρήνα και δίνει μία σταθερή τιμή για κάθε χαρακτηριστικό μεταξύ 0 και 1. Στο κομμάτι της απόφασης χρησιμοποιούμε το DAG αντί για τη μέθοδο ένας εναντίον ενός (*One against One-OAO*) ή αντί ένας εναντίον όλων (*One against All-OAA*), το οποίο στον ίδιο χρόνο πετυχαίνει μεγαλύτερη ακρίβεια.

3.6 Μοντέλο αναγνώρισης αλλαγών σε διδιάστατη διαδικασία που βασίζεται στον αναγνωριστή μοτίβων $LS-SVM$

Προτείνεται (Qiang et al. 2010) επίσης ένα μοντέλο που βασίζεται στις μηχανές διανυσματικής υποστήριξης ελάχιστων τετραγώνων (*Least Squares Support Vector Machines-*$LS-SVM$) για διδιάστατες διαδικασίες, το οποίο βρίσκει μη φυσικά μοτίβα και βοηθά στην αναγνώριση των μη φυσικών μεταβλητών, όταν χρησιμοποιούνται τα *Shewhart* πολυμεταβλητά διαγράμματα ελέγχου που βασίζονται στο *Hotelling's* T^2 διάγραμμα. Αυτό το μοντέλο είναι ένα καλό συμπλήρωμα στα *Shewhart* διαγράμματα ελέγχου. Η απόδοση του εκτιμάται από τον υπολογισμό της ακρίβειας της ταξινόμησης του $LS-SVM$ αναγνωριστή μοτίβων.

Στο Σχήμα 3.5 φαίνεται σχηματικά η κατασκευή του μοντέλου.

Το μοντέλο περιλαμβάνει δύο τμήματα, το πρώτο χρησιμοποιεί το στατιστικό *Hotelling* T^2 για να ανιχνεύσει όλα τα εκτός ελέγχου σήματα στη διδιάστατη διεργασία. Όταν το στατιστικό *Hotelling* T^2 προειδοποιήσει, ο αναγνωριστής μοτίβων $LS-SVM$ του δεύτερου τμήματος μοντελοποιεί το πρόβλημα επιλογής



Σχήμα 3.5: Μοντέλο διάγνωσης διεργασίας σε T^2 διάγραμμα και $LS-SVM$ αναγνωριστή μοτίβων

αιτίου σαν πρόβλημα αναγνώρισης μοτίβου, δηλαδή το αφύσικο σήμα μπορεί να ταξινομηθεί σε διαφορετικές κλάσεις μοτίβων. Υπάρχουν τρεις διακεκριμένες κλάσεις μοτίβων:

1. η πρώτη μεταβλητή είναι εκτός ελέγχου,
2. η δεύτερη μεταβλητή είναι εκτός ελέγχου,
3. και η πρώτη και η δεύτερη μεταβλητή είναι εκτός ελέγχου.

Είναι σαφές ότι το αφύσικο σήμα πρέπει να περιέχεται σε μία από αυτές τις περιπτώσεις. Έτσι ο $LS-SVM$ αναγνωριστής μοτίβου μπορεί να εφοδιάσει τη μέθοδο με την ακριβή πληροφορία για τη μη-φυσική μεταβλητή έτσι ώστε να γίνουν οι κατάλληλες ενέργειες στη συνέχεια. Το μοντέλο θεωρεί ότι για μία p-διάστατη κατασκευαστική διεργασία με p μεταβλητές, ο μέσος κάθε μεταβλητής έχει δύο καταστάσεις: την κανονική και την μη-κανονική. Συνεπώς υπάρχουν 2^p πιθανές καταστάσεις. Δεν υπάρχει αμφιβολία οτι μόνο μία στις 2^p καταστάσεις είναι η κανονική. Όταν το στατιστικό T^2 εντοπίσει κάποια ανωμαλία στη διαδικασία, υπάρχουν (2^p-1) διαφορετικές μη φυσικές καταστάσεις. Το μοντέλο καθορίζει τις (2^p-1) πιθανές μη φυσικές καταστάσεις σαν (2^p-1) μη φυσικά μοτίβα που πρέπει να αναγνωριστούν. Όταν το T^2 στατιστικό δώσει κάποιο σήμα, το μοντέλο χρησιμοποιεί τον εκπαιδευμένο $LS-SVM$ αναγνωριστή μοτίβων για να ταξινομήσει τα αφύσικα μοτίβα του διανύσματος του μέσου της διαδικασίας και επιπλέον αναγνωρίζει τα αφύσικα συστατικά του διανύσματος του μέσου.

3.6.1 Ο αναγνωριστής μοτίβων $LS-SVM$

Το SVM έχει μεγάλη ικανότητα στην ταξινόμηση δεδομένων, όταν το μέγεθος των δειγμάτων είναι μικρό. Οι μηχανές διανυσμάτων υποστήριξης ελαχίστων τετραγώνων $LS-SVM$ είναι μία επέκταση του SVM (Suykens and Vanderwalle, 1999). Το SVM είναι εκπαιδευμένο να λύνει προβλήματα τετραγωνικής βελτιστοποίησης, ενώ το $LS-SVM$ είναι εκπαιδευμένο να λύνει γραμμικές εξισώσεις, οπότε πετυχαίνει πιο μεγάλη ταχύτητα στους υπολογισμούς. Παρακάτω δίνεται μία περιγραφή της μεθόδου $LS-SVM$.

Έστω x_i ένα σύνολο δεδομένων και y_i οι αντίστοιχες ομάδες με $x_i \in R^n$, $y_i \in R$ και $i = 1, ..., N$. Η τεχνική $LS-SVM$ διατυπώνει το πρόβλημα βελτιστοποίησης ως:

$$\begin{cases} \min & J(\omega,\xi) = \frac{1}{2}\omega^T\omega + \frac{1}{2}\gamma\sum_{i=1}^{N}\xi^2 \\ \text{δεδομένου ότι} & y_i = \omega^T\phi(\chi_i) + b + \xi_i \end{cases}, \tag{3.4}$$

όπου $\phi(.)$ είναι μια μη γραμμική συνάρτηση αντιστοίχισης. Έτσι το παραπάνω πρόβλημα βελτιστοποίησης μετατρέπεται στο πρόβλημα λύσης της παρακάτω γραμμικής εξίσωσης:

$$\begin{pmatrix} \Omega + \gamma^{-1}I & l_N \\ l_N^T & 0 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} y \\ 0 \end{pmatrix}, \tag{3.5}$$

όπου $l_N = [1, 1, \ldots, 1]^T, \alpha = [\alpha_1, \alpha_2, \ldots, \alpha_N]^T, y = [y_1, y_2, \ldots, y_N]^T$. Ω είναι ένας τετραγωνικός πίνακας με στοιχεία Ω_{ij}=$\psi(x_i, y_i)$=$\phi(x_i)^T\phi(x_i)$, i, j=1,...,N. $\psi(.)$ είναι η συνάρτηση πυρήνα που ικανοποιεί το θεώρημα του $Mercer$. Τα a, b είναι οι μοναδίκες λύσεις της εξίσωσης (3.5) και το γραμμικό μοντέλο παλινδρόμησης παρουσιάζεται και ως εξής:

$$y(x) = \sum_{i=1}^{N}\alpha_i\psi(x, x_i) + b \tag{3.6}$$

Στη συγκεκριμένη περίπτωση ως ψ επιλέγουμε τη $Gaussian$ συνάρτηση πυρήνα, οπότε το ψ έχει τη μορφή:

$$\psi(x, y) = \exp\left(\frac{-\parallel x - y \parallel^2}{2\sigma^2}\right) \tag{3.7}$$

Το τελικό $LS-SVM$ μοντέλο για την εκτίμηση συνάρτησης

$$f(x) = \sum_{i=1}^{N} y_i\alpha_i\psi(x, x_i) \tag{3.8}$$

Και η τελική συνάρτηση απόφασης είναι $sign[f(x)]$, όπου

$$sign[f(x)] = \begin{cases} 1, & x > 0 \\ -1, & x < 0 \end{cases} .$$

Έτσι το μέσο διάνυσμα μπορεί να χρησιμοποιηθεί σαν είσοδος (x_i) και τα φυσικά και μη μοτίβα επισημαίνονται σαν y_i.

4 SVM και ανίχνευση λαθών

4.1 Τα σφάλματα στο σύστημα μέτρησης

Η ζήτηση σε ότι αφορά την αποτελεσματικότητα της παραγωγής, την ποιότητα των προϊόντων, τα επίπεδα ασφαλείας και την προστασία του περιβάλλοντος συνεχώς αυξάνεται στις βιομηχανίες παραγωγής. Ο τρόπος για να ικανοποιήσουμε τη ζήτηση είναι να παράγουμε όλο και πιο πολύπλοκα αυτόματα συστήματα ελέγχου που απαιτεί περισσότερες μεταβλητές για μετρήσεις και πιο εξελιγμένα συστήματα μέτρησης. Οι ποιοτικές και αξιόπιστες μετρήσεις είναι η βάση για τον ποιοτικό έλεγχο διεργασιών. Μία βλάβη στον εξοπλισμό της διεργασίας χειροτερεύει την παραγωγή και ακόμα μπορεί να προκαλέσει τη διακοπή της παραγωγής και να επιφέρει πρόσθετες δαπάνες. Ο έλεγχος διεργασιών εξαρτάται σε μεγάλο βαθμό από την ποιότητα των δεδομένων, οπότε είναι κρίσιμο να μετρηθούν όσο περισσότερες μεταβλητές είναι δυνατόν και να ανακαλυφθούν πιο εξελιγμένα συστήματα μέτρησης. Παρακάτω αναλύεται η αυτόματη ανίχνευση λαθών και η αναγνώριση του εξοπλισμού μετρήσεων της διαδικασίας.

Η διακοπή της παραγωγής μπορεί να προκληθεί από διάφορες μη φυσικές καταστάσεις όπως:

>Βλάβη στον εξοπλισμό

- βλάβη του εργοστασίου,
- βλάβη στον εξοπλισμό μετρήσεων,
- βλάβη στο σύστημα επικοινωνίας.

>Μεγάλες διαταραχές

Σε πολλές περιπτώσεις, οι βλάβες στον εξοπλισμό επιδρούν αρνητικά στην παραγωγή και στην ποιότητα του προϊόντος. Τα πιο επικίνδυνα σφάλματα είναι αυτά στον εξοπλισμό των μετρήσεων λόγω του ότι το γενικό σύστημα του ελέγχου του συστήματος παραγωγής στηρίζεται στις μετρήσεις. Η δράση του ελέγχου βασίζεται στα δεδομένα που προέρχονται από ελαττωματικό αισθητήρα, που είναι στην καλύτερη περίπτωση ανεπαρκή και στη χειρότερη επικίνδυνα. Επίσης είναι σημαντικό να ελέγχεται το νόημα των συλλεγόμενων δεδομένων μέτρησης, πριν αυτά διαβιβαστούν στο αύτοματο σύστημα ελέγχου για περαιτέρω επεξεργασία.

Στα εργοστάσια υπάρχουν πολλές μεταβλητές που μετρούνται από αισθητήρες και καταγράφονται στη βάση δεδομένων της διεργασίας, έτσι η ποσότητα

των διαθέσιμων δεδομένων είναι μεγάλη. Κάτω από κανονικές συνθήκες λειτουργίας αυτές οι μεταβλητές είναι ισχυρά συνδεδεμένες εξαιτίας φυσικών και χημικών αρχών. Αυτές οι σχέσεις μπορούν να μοντελοποιηθούν, και το μοντέλο που προκύπτει μπορεί να χρησιμοποιηθεί για να ελέγξει νέα δεδομένα για να ανιχνεύσει μη φυσικές καταστάσεις. Αφού το μοντέλο κατασκευάζεται από τα δεδομένα του εργοστασίου, είναι σημαντικό να επιλεχθούν οι κατάλληλες τεχνικές μοντελοποίησης. Οι τεχνικές της πολυμεταβλητής ανάλυσης μπορούν να χρησιμοποιηθούν για την ανάλυση ισχυρά συσχετισμένων δεδομένων και την παρακολούθηση της διεργασίας. Αυτή η προσέγγιση ονομάζεται παρακολούθηση πολυμεταβλητής στατιστικής διεργασίας ($Multivariate\ Statistical\ Process\ Monitoring - MSPM$) και έχει πολλές εφαρμογές σε διάφορες βιομηχανικές διεργασίες. Βασικά η $MSPM$ αποτελείται από δύο κύρια μέρη: την ανίχνευση λαθών και την αναγνώριση τους. Η απόδοσή του εξαρτάται από το πόσο καλά το μοντέλο περιγράφει την σχέση μεταξύ των μεταβλητών. Η πιο γνωστή μέθοδος για την μοντελοποίηση αυτών των σχέσεων είναι η PCA. Όμως η PCA θεωρεί τις σχέσεις μεταξύ των μεταβλητών γραμμικές και $Gaussian$ λανθάνουσες μεταβλητές. Έτσι μπορεί να είναι μη αποτελεσματική όταν αντιμετωπίζει βιομηχανικές διεργασίες που έχουν συνήθως μη γραμμικές και έχουν $non - Gaussian$ βασικές μεταβλητές. Πρόσφατα εμφανίστηκαν καινούριες τεχνικές όπως η ανάλυση ανεξάρτητων συνιστωσών ($Independent\ Components\ Analysis - ICA$) και διάφορες μέθοδοι πυρήνα. Αυτές οι μέθοδοι μπορούν μερικές φορές να εκμεταλλευτούν τα δεδομένα εργοστασίου με πιο αποτελεσματικό τρόπο από τη μέθοδο PCA.

Η μέθοδος PCA:

Η αφετηρία για κάθε πολυμεταβλητή ανάλυση είναι ο πίνακας δεδομένων $\mathbf{X}$, του οποίου τα στοιχεία είναι τα μετρήσιμα δεδομένα που λαμβάνονται από τις μετρήσεις που γίνονται στη διεργασία. Οι m στήλες του πίνακα $\mathbf{X}$ καλούνται αντικείμενα και οι n γραμμές αντιπροσωπεύουν τη διάσταση του χώρου εισόδου. Η PCA είναι μία μέθοδος για την προβολή-σχεδίαση του υψηλής διάστασης, συσχετισμένου χώρου εισόδου σε έναν κατάλληλο υπόχωρο μικρότερης διάστασης. Η PCA ψάχνει στο χώρο είσοδου δεδομένων για τις κατευθύνσεις των μεταβολών των μεγάλων δεδομένων, υπό την προϋπόθεση ότι αυτές οι διευθύνσεις είναι ορθογώνιες και τις χρησιμοποιούν σαν κύριο άξονα ενός νέου ισότιμου συστήματος, στο οποίο ο χώρος εισόδου είναι προβλεπόμενος. Εκ τούτου η PCA μεταφέρει συσχετισμένες μεταβλητές σε μία ομάδα καινούριων μη συσχετισμένων μεταβλητών, οι οποίες καλούνται κύριες συνιστώσες. Εκτός από το ότι είναι ασυσχέτιστες, οι νέες μεταβλητές είναι ταξινομημένες ανάλογα με το μέγεθος της μεταβολής των δεδομένων που περιγράφουν. Οι μεγαλύτερες

κύριες συνιστώσες χρησιμοποιούνται για την εκτίμηση μεταβλητών της διεργασίας και οι μικρότερες χρησιμοποιούνται για την ανίχνευση λαθών και την αναγνώρισή τους.

Η μέθοδος *ICA*:

Ο στόχος της ICA είναι να αναλύσει δεδομένα σε γραμμικούς συνδυασμούς σε στατιστικά ανεξάρτητες συνιστώσες. Οι μετρήσιμες μεταβλητές της διεργασίας είναι συνήθως ένας συνδυασμός από ανεξάρτητες μεταβλητές που δεν είναι άμεσα μετρήσιμες, οπότε η παρακολούθηση της διεργασίας που βασίζεται στην ICA είναι καλύτερη από αυτή που βασίζεται στην PCA. Η μείωση της διάστασης στην ICA βασίζεται στην πεποίθηση ότι οι μετρήσιμες μεταβλητές είναι μίξεις από μικρό αριθμό ανεξάρτητων μεταβλητών. Σε αντίθεση με τη μέθοδο της PCA, η επιλογή των ανεξάρτητων συνιστωσών δεν είναι τετριμμένη. Ενώ ο αυθεντικός αλγόριθμος ICA έχει αρκετά μειονεκτήματα, οι επεκτάσεις της μεθόδου αυτής επαυξάνουν την ICA που βασίζεται στην παρακολούθηση διεργασιών. Στην αυθεντική ICA ο αριθμός των ICs (ανεξάρτητων συνιστωσών) είναι ίσος με τον αριθμό των μετρήσιμων μεταβλητών που σημαίνει ότι εξάγονται κάποιες ασήμαντες ICs από τα μετρήσιμα δεδομένα. Εκτός αυτού, οι ICs δεν είναι ταξινομημένες βάσει της σημαντικότητας όπως είναι στη μέθοδο PCA. Στα τροποποιημένα ICA προτείνεται να εξάγονται μόνο λίγες κυρίαρχες ICs, καθορίζεται η σειρά των ICs και δίνεται μία συνεπής λύση.

Ανίχνευση λαθών:

Η ανίχνευση λαθών καθορίζει αν έχει προκύψει ένα σφάλμα. Οι πολυμεταβλητές στατιστικές τεχνικές μπορούν να χρησιμοποιηθούν για να ανιχνεύσουν τις παρακάτω μη φυσικές καταστάσεις αισθητήρα:

- Οι μετρήσεις παίρνουν μη φυσικές τιμές, συνήθως λόγω ενός μεγάλου σφάλματος αισθητήρα.
- Πολλοί αισθητήρες αποκλίνουν από κανονικούς συσχετισμούς.
- Η διαδικασία που παρακολουθείται υφίσταται παροδικές διακυμάνσεις.

Για την ανίχνευση των σφαλμάτων, αναπτύσσεται το μοντέλο PCA, που βασίζεται σε δεδομένα της διεργασίας όταν λειτουργεί κανονικά, και στη συνέχεια ελέγχει τα καινούρια δεδομένα των μετρήσεων. Οι διαφορές μεταξύ των καινούριων δεδομένων μετρήσεων και των προβολών τους στο μοντέλο, δηλαδή τα

υπόλοιπα υποβάλλονται σε ένα είδος στατιστικού ελέγχου για να καθοριστεί αν είναι σημαντικά. Συνήθως το στατιστικό Q, το οποίο καλείται επίσης τετραγωνικό σφάλμα πρόβλεψης (*squared prediction error-SPE*), και το στατιστικό *Hotelling* (T^2) χρησιμοποιούνται για το συμβολισμό της μεταβλητότητας στον υπόχωρο υπολοίπων και στον υπόχωρο κύριων συνιστωσών.

Το στατιστικό Q δείχνει πόσο καλά προσαρμόζεται ένα καινούριο δείγμα στο μοντέλο PCA που έχει φτιαχτεί από τα προηγούμενα δεδομένα μέτρησης. Τα υπόλοιπα είναι ένα μέτρο της διαφοράς μεταξύ του δείγματος και των προβολών στις l κύριες συνιστώσες που δίνει το μοντέλο. Ο τύπος του υπολοίπου (r_i) για το x_i δίνεται από τον τύπο:

$$\mathbf{r_i} = \mathbf{x_i} - \hat{\mathbf{x}}_\mathbf{i} = \mathbf{x_i}(\mathbf{I} - \mathbf{P_l}\mathbf{P_l^T}) \tag{4.1}$$

και το μέγεθός του υπολογίζεται από τον τύπο:

$$Q = \|\mathbf{r_i}\| = \mathbf{r_i}\mathbf{r_i^T} = \mathbf{x_i}(\mathbf{I} - \mathbf{P_l}\mathbf{P_l^T})\mathbf{x_i^T} \tag{4.2}$$

και αναπαριστά το πόσο καλή είναι η προσαρμογή του καινούριου δείγματος στο μοντέλο P_l σαν $scalar$. Τα όρια εμπιστοσύνης μπορούν να υπολογιστούν από το υπόλοιπο του μοντέλου Q, δεδομένου ότι όλες οι ιδιοτιμές του πίνακα διασποράς είναι γνωστές:

$$Q_a = \Theta_1[\frac{c_a\sqrt{2\Theta_2 h_0^2}}{\Theta_1} + 1 + \frac{\Theta_2 h_0(h_0-1)}{\Theta_1^2}]\frac{1}{h_0} \tag{4.3}$$

όπου

$$\Theta_i = \sum_{j=k+1}^{n} \lambda_j^i, i = 1, 2, 3,$$
$$h_0 = 1 - \frac{2\Theta_1\Theta_3}{3\Theta_2^2},$$

όπου Q_α είναι το άνω όριο εμπιστοσύνης για το μοντέλο υπολοίπων Q με επίπεδο εμπιστοσύνης α και c_α είναι οι αντίστοιχες αποκλίσεις.

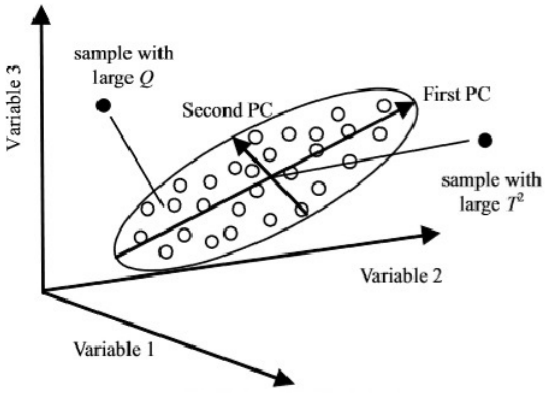
Το στατιστικό του *Hotelling*-T^2 παρέχει μία ένδειξη για τη μη φυσική μεταβλητότητα στον κανονικό υπόχωρο. Η τιμή του T^2 για ένα δείγμα είναι ίση με το άθροισμα των τεραγώνων των προσαρμοσμένων αποτελεσμάτων για κάθε μία από τις κύριες συνιστώσες του μοντέλου:

$$T^2 = \sum_{i=1}^{l}(\frac{t_i}{\lambda_i})^2. \tag{4.4}$$

Το T^2 αναπαριστά το τετραγωνικό μήκος της προβολής του τρέχοντος δείγματος στο χώρο που διευρύνεται (*spanned*) από το μοντέλο PCA των δεδομένων. Είναι μία ένδειξη για το πόσο διαφέρει η εκτίμηση του δείγματος από το PCA από το πολυμεταβλητό μέσο των δεδομένων, δηλαδή των κύριων συνιστωσών. Επομένως, αν το δείγμα δεν έχει μία φυσιολογική τιμή του T^2 αλλά η τιμή του Q είναι εντός των ορίων δεν είναι απαραίτητο να υπάρχει σφάλμα, μπορεί να ευθύνεται η αλλαγή στην περιοχή λειτουργίας. Τα στατιστικά όρια εμπιστοσύνης για τις τιμές του T^2 υπολογίζονται σύμφωνα με την F-κατανομή:

$$T^2_{l,n,\alpha} = \frac{l(n-1)}{n-l} F_{l,n-l,\alpha}, \tag{4.5}$$

όπου n είναι ο αριθμός των δειγμάτων στο σύνολο των δεδομένων που χρησιμοποιείται για το μοντέλο PCA, l είναι ο αριθμός των κύριων συνιστωσών που έχουν διατηρηθεί (*retained*) και α είναι η παράμετρος της τυπικής κανονικής απόκλισης.



Σχήμα 4.1: Προβολή των δεδομένων για δύο PCs

Στο Σχήμα 4.1 φαίνονται τα όρια του T^2 που καθορίζουν μία έλλειψη εντός της οποίας τα δεδομένα υποτίθεται ότι είναι κανονικά.

Αναγνώριση των σφαλμάτων:

Αφού ανιχνευθεί ένα σφάλμα, είναι σημαντικό να διαγνωστεί η αιτία της αποτυχίας. Η πιο δημοφιλής προσέγγιση είναι η χρήση διαγραμμάτων συνεισφοράς (*contribution plots*). Ένας εσφαλμένος αισθητήρας συνήθως καταστρέφει τη κανονική συσχέτιση με τους υπόλοιπους αισθητήρες. Τα χαρακτηριστικά μπορούν να χρησιμοποιηθούν για την αναγνώριση του εσφαλμένου αισθητήρα αφού

ανιχνευθεί μία μη φυσική κατάσταση. Το διάγραμμα συνεισφοράς χρησιμοποιεί το υπόλοιπο για κάθε αισθητήρα σε κάθε δείγμα για να αναγνωρίσει τους αισθητήρες που σχετίζονται με το σφάλμα που ανιχνεύθηκε. Ο αισθητήρας με τη μεγαλύτερη τιμή σφάλματος θεωρείται ότι έιναι εσφαλμένος, αφού έχει τη μεγαλύτερη συνεισφορά στο τετραγωνικό σφάλμα πρόβλεψης. Συνήθως είναι δυνατό να ξανακατασκευαστούν οι εσφαλμένοι αισθητήρες που βασίζονται στα πολυμεταβλητά στατιστικά μοντέλα για να διατηρείται ο έλεγχος της διεργασίας και η βελτιστοποίηση. Ωστόσο είναι δύσκολο να αναγνωριστούν οι εσφαλμένοι αισθητήρες από τα διαγράμματα συνεισφοράς, όταν προκύπτει σφάλμα από δύο αισθητήρες ταυτόχρονα. Τα αποτελέσματα της έρευνας (Slišković et al. (2012)) δείχνουν ότι η παρακολούθηση που βασίζεται στην ICA είναι πιο ευαίσθητη από την PCA μέθοδο. Απ΄ την άλλη, η μέθοδος PCA είναι πιο απλή από την ICA όσο αφορά το υπολογιστικό κόστος και τη διαδικασία βελτιστοποίησης που πρέπει να γίνει στη μελέτη όταν τέτοιες μέθοδοι εκτελούνται στην επίβλεψη ενός εργοστασίου και στο σύστημα ελέγχου. Η τιμή της μεταβλητής του εσφαλμένου αισθητήρα μπορεί να εκτιμηθεί από τους υπόλοιπους αισθητήρες με διάφορες τεχνικές ανίχνευσης.

4.2 Η χρήση μιας υπολογιστικά έξυπνης υβριδικής προσέγγισης για την αναγνώριση των λαθών στη διακύμανση των αλλαγών σε μία κατασκευαστική διαδικασία.

Τα διαγράμματα του ΣΕΔ είναι αποτελεσματικά στην παρακολούθηση μίας διαδικασίας. Στην περίπτωση της μονομεταβλητής διεργασίας, δεν είναι δύσκολο να καθοριστεί το προσδιορισμένο αίτιο λόγω του ότι στο μονομεταβλητό SPC διάγραμμα παρακολουθείται ένα μεμονωμένο ποιοτικό χαρακτηριστικό. Το εκτός ελέγχου σήμα συνεπάγεται ότι η μοναδική μεταβλητή της διεργασίας έχει κάποιο πρόβλημα και το προσωπικό της διαδικασίας πρέπει να δώσει προσοχή στη μεταβλητή έτσι ώστε να διορθωθεί το πρόβλημα. Στην περίπτωση όμως που χρησιμοποιείται το πολυμεταβλητό SPC διάγραμμα για την παρακολούθησης μιας πολυμεταβλητής διαδικασίας είναι δύσκολο να καθοριστεί ποιο ποιοτικό χαρακτηριστικό είναι υπεύθυνο για το σφάλμα που προκύπτει. Παρακάτω προτείνεται ένα μοντέλο υβριδικής ταξινόμησης ($hybrid\ classification$) που αναγνωρίζει τα ποιοτικά χαρακτηριστικά που είναι υπεύθυνα για τις αλλαγές που προκύπτουν σε μία πολυμεταβλητή διεργασία. Ο μηχανισμός που προτείνεται περιλαμβάνει την παραγωγή μικτών γενών ($hybridization$) του τεχνητού νευρωνικού δικτύου ($Artificial\ Neural\ Network$-$ANN$) και την ανάλυση της

διακύμανσης (*analysis of variance-ANOVA*).

Όταν ο $MSPC$ δώσει ένα εκτός ελέγχου σήμα, υποδεικνύει ότι έχει προκύψει σφάλμα στη διεργασία, αλλά δεν αναφέρεται ποιο ποιοτικό χαρακτηριστικό ή ποια ομάδα χαρακτηριστικών ευθύνονται για αυτό το σφάλμα. Πριν γίνουν οι διορθωτικές ενέργειες, το προσωπικό της διαδικασίας πρέπει πρώτα να αναγνωρίσει ή να καθορίσει τις ποιοτικές μεταβλητές που ευθύνονται για το σφάλμα. Αυτή η αναγνώριση δεν είναι εύκολη. Όσο περισσότερες μεταβλητές περιέχονται σε μία διεργασία, τόσο αυξάνεται ο βαθμός δυσκολίας για την αναγνώριση.

Όταν προκύψει σήμα εκτός ελέγχου στον $MSPC$, το προσωπικό της διεργασίας πρέπει να ψάξει τα πιθανά αίτια του προβλήματος. Ο καθορισμός αυτών των αιτίων είναι πολύ δύσκολο κομμάτι και πολλές έρευνες ασχολούνται με αυτό. Οι γραφικές προσεγγίσεις χρησιμοποιούνται για να βοηθήσουν στον καθορισμό των ποιοτικών χαρακτηριστικών που είναι υπεύθυνα για τα σφάλματα σε μία διαδικασία, όμως τα αποτελέσματά τους είναι ανιαρά και υποκειμενικά. Οι μέθοδοι στατιστικής αποσύνθεσης χρησιμοποιούνται για να ερμηνεύσουν τα αίτια ενός σήματος σε ένα SPC. Για παράδειγμα, οι Mason et al. (1995) πρότειναν μία χρήσιμη προσέγγιση στην αποσύνθεση του T^2 στατιστικού σε ανεξάρτητα κομμάτια, καθένα από τα οποία αντικατοπτρίζει τη συνεισφορά μίας μεμονωμένης ποιοτικής μεταβλητής. Η PCA ερευνάται για την αποτελεσματικότητα της στο να καθορίσει τα ζητούμενα ποιοτικά χαρακτηριστικά σε μία πολυμεταβλητή διεργασία. Ωστόσο η PCA προσέγγιση υποστηρίζει ότι η διάσταση των δεδομένων δεν μειώνεται αποτελεσματικά με τη γραμμική μεταφορά. Επίσης το πρόβλημα του PCA είναι ότι η μεγιστοποίηση των μεταβλητών δεν μεγιστοποιεί πάντα την πληροφορία.

Πρόσφατα κάποιες υπολογιστικά έξυπνες προσεγγίσεις, χρησιμοποιήθηκαν για να καθορίσουν ποιο ποιοτικό χαρακτηριστικό είναι υπεύθυνο για το σήμα στο SPC. Και οι δύο έρευνες (Aparisi et al. (2006), Shao and Hsu (2009)) κατέληξαν στο συμπέρασμα ότι η απόδοση αυτών των υπολογιστικά έξυπνων μεθόδων είναι καλύτερη από την προσέγγιση μέσω της αποσύνθεσης που αναφέρθηκε παραπάνω. Πρόσθετα ανακαλύφθηκε (Niaki και Abbasi (2005)) ένα μοντέλο δύο επιπέδων για τον καθορισμό της πηγής των σημάτων εκτός ελέγχου.

Ένα μοντέλο που βασίζεται στο ANN προτείνεται για να προσδιορίσει και ποσοτικοποιήσει τις αλλαγές στο μέσο για μία διδιάστατη διεργασία. Ενώ οι περισσότερες μελέτες χρησιμοποιούν την αλλαγή στο μέσο της διαδικασίας σαν υπαίτιο, στη συγκεκριμένη μελέτη σαν σφάλμα θεωρείται η αλλαγή στη μεταβλητότητα της διαδικασίας. Πρόκειται για μία πολυμεταβλητή διαδικασία με 7 ποιοτικά χαρακτηριστικα τα οποία παρακολουθούνται από το $|S|$, ένα διάγραμμα $MSPC$. Επιπρόσθετα θεωρείται ότι ο πίνακας διασποράς αλλάζει από Σ_0 σε Σ_1, όταν προκύπτει το σφάλμα. Υπάρχουν 29 εισαγόμενες μεταβλητές που

πρέπει να εξεταστούν και δεν είναι πρακτικό να χρησιμοποιηθούν και οι 29 σαν δεδομένα εισαγωγής στον ταξινομητή ANN. Συνεπώς η μελέτη χρησιμοποιεί μία υβριδική τεχνική για να επιλέγει λιγότερες αλλά πιο σημαντικές επεξηγηματικές μεταβλητές. Αυτή είναι η αρχική φάση για να φτιάξουμε την προτεινόμενη μέθοδο. Στη δεύτερη φάση οι μεταβλητές που επιλέχθηκαν γίνονται τα δεδομένα εισαγωγής για τα ANN μοντέλα.

4.3 Περιγραφή του μοντέλου

Έστω ότι η πολυμεταβλητή διεργασία είναι αρχικά εντός ελέγχου, και το δείγμα των παρατηρήσεων λαμβάνεται απο μία άγνωστη κατανομή $F(\boldsymbol{\mu},\boldsymbol{\Sigma})$ με γνωστό μέσο $\boldsymbol{\mu}$ και πίνακα διασποράς $\boldsymbol{\Sigma}_0$. Μετά από την εμφάνιση ενός σφάλματος στην αλλαγή της μεταβλητότητας σε ένα συγκεκριμένο χρόνο, η συγκεκριμένη μελέτη θεωρεί ότι ο πίνακας διασποράς της διεργασίας αλλάζει από $\boldsymbol{\Sigma}_0$ σε $\boldsymbol{\Sigma}_1$:

$$\boldsymbol{\Sigma_0} = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \dots & \sigma_{1,j} & \dots & \sigma_{1,p} \\ \sigma_{2,1} & \sigma_{2,2} & \dots & \vdots & \dots & \sigma_{2,p} \\ \vdots & \vdots & \ddots & \vdots & \dots & \vdots \\ \sigma_{i,1} & & & \sigma_{i,j} & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{p,1} & \sigma_{p,2} & \dots & \sigma_{p,j} & \dots & \sigma_{p,p} \end{bmatrix}_{p\times p} \tag{4.6}$$

$$\boldsymbol{\Sigma_1} = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \dots & \vartheta\sigma_{1,j} & \sigma_{1,j+1} & \dots & \sigma_{1,p} \\ \sigma_{2,1} & \sigma_{2,2} & \dots & \vartheta\sigma_{2,j} & \sigma_{2,j+1} & \dots & \sigma_{2,p} \\ \vdots & \vdots & \ddots & \vdots & \dots & \vdots & \\ \vartheta\sigma_{i,1} & \vartheta\sigma_{i,2} & & \vartheta^2\sigma_{i,j} & \vartheta\sigma_{i,j+1} & \dots & \vartheta\sigma_{i,p} \\ \sigma_{i+1,1} & \sigma_{i+1,2} & \dots & \vartheta\sigma_{i+1,j} & \sigma_{i+1,j+1} & \dots & \sigma_{i+1,p} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \\ \sigma_{p,1} & \sigma_{p,2} & \dots & \vartheta\sigma_{p,j} & \sigma_{p,j+1} & \dots & \sigma_{1,p} \end{bmatrix}_{p\times p}, \tag{4.7}$$

όπου το ϑ είναι ο διογκωμένος λόγος ($inflated\,ratio$). Ο πίνακας διακύμανσης - συνδιακύμανσης για την υποομάδα i ορίζεται ως εξής:

$$S_i = \frac{1}{n-1}\sum_{j=1}^{n}(\underset{\sim}{X}_{i,j} - \underset{\sim}{\bar{X}}_i)(\underset{\sim}{X}_{i,j} - \underset{\sim}{\bar{X}}_i)'$$

$$= \begin{bmatrix} S_{i,1,1} & S_{i,1,2} & \dots & S_{i,1,p} \\ S_{i,2,1} & S_{i,2,2} & \dots & S_{i,2,p} \\ \vdots & \vdots & \vdots & \vdots \\ S_{i,p,1} & S_{i,p,2} & \dots & S_{i,p,p} \end{bmatrix}_{p\times p}.$$

Επιπρόσθετα το δείγμα των γενικευμένων διακυμάνσεων $|S_i|$, ι=1,2,... και τα ακόλουθα όρια (UCL, LCL) χρησιμοποιούνται για να παρακολουθούν τη μεταβλητότητα στις αλλαγές σε μία πολυμεταβλητή διεργασία.

$$UCL = |\mathbf{\Sigma_0}|(b_1 + 3\sqrt{b_2}) \tag{4.8}$$

$$LCL = max(0, |\mathbf{\Sigma_0}|(b_1 - 3\sqrt{b_2})), \tag{4.9}$$

όπου το $|\mathbf{\Sigma_0}|$ είναι η ορίζουσα του $\mathbf{\Sigma_0}$ και

$$b_1 = \frac{1}{(n-1)^p}\prod_{i=1}^{p}(n-i), \tag{4.10}$$

$$b_2 = \frac{1}{(n-1)^{2p}}\prod_{i=1}^{p}(n-i)(\prod_{i=1}^{p}(n-i+2) - \prod_{i=1}^{p}(n-i)). \tag{4.11}$$

Η προτεινόμενη υβριδική προσέγγιση συνδυάζει το σκελετό του $ANOVA$ με το ANN. Πρώτα εφαρμόζεται ένας $one-way\ ANOVA$ έλεγχος για να επιλέξει τις σημαντικές μεταβλητές που έχουν μεγάλη επιρροή. Έπειτα οι επιλεγμένες σημαντικές μεταβλητές εισάγονται στον ταξινομητή ANN.

Το μοντέλο *ANOVA* :

Ο λόγος που χρησιμοποιούμαι το $one-way\ ANOVA$ στην αρχική φάση είναι για να προσδιορίσουμε ποια δεδομένα από τις εντός και εκτός ελέγχου ομάδες έχουν ίδιο μέσο, δηλαδή να προσδιορίσουμε αν τα μετρήσιμα χαρακτηριστικά από τις ομάδες εντός και εκτός ελέγχου είναι όντως διαφορετικά. Λόγω του ότι ο πίνακας S_i είναι συμμετρικός, εξετάζονται μόνο τα στοιχεία που βρίσκονται πάνω από τη διαγώνιο. Για να απλοποιηθεί ο συμβολισμός αντικαθιστούμε:

$$Y_{i,1} = S_{i,1,1}, Y_{i,2} = S_{i,1,2}, ..., Y_{i,p} = S_{i,1,p}$$

$$Y_{i,p+1} = S_{i,2,2}, Y_{i,p+2} = S_{i,2,3}, ..., Y_{i,N-1} = S_{i,p,p}, Y_{i,N} = |S|,$$

όπου N=1+$p(p+1)/2$. Το $Y_{i,j,k,l}$ είναι η l-οστή παρατήρηση του k-οστού επιπέδου του παράγοντα για τη μεταβλητή $Y_{i,j}$ που αναφέρεται παραπάνω, $(i = 1, 2, ..., T; j = 1, 2, ..., N; k = 1, 2; l = 1, 2, ..., n_{i,j,k})$ Έστω ότι $\mu_{i,j}$ και $\tau_{i,j,k}$ είναι αντίστοιχα ο γενικός μέσος και η επίδραση της αγωγών $(treatment\ effect)$. Αναλόγως η γραμμική εξίσωση για το μοντέλο $one - way\ ANOVA$ είναι:

$$Y_{i,j,k,l} = \mu_{i,j} + \tau_{i,j,k} + \epsilon_{i,j,k,l}, \tag{4.12}$$

όπου $i = 1, 2, ..., T; j = 1, 2, ..., N; k = 1, 2; l = 1, 2, ..., n_{i,j,k}$ Για να αναγνωρίσουμε τις σημαντικές μεταβλητές, χρησιμοποιείται ένα F στατιστικό που ελέγχει τις διαφορές μεταξύ των ομάδων εντός και εκτός ελέγχου. Αυτές οι σημαντικές μεταβλητές επιλέγονται σε αυτή τη φάση και έπειτα αντικαθίστονται στο ANN για να κατασκευαστεί το υβριδικό μοντέλο.

Το μοντέλο *ANN:*

Η χρήση του ANN για την πρόβλεψη και μοντελοποίηση έχει μεγάλο ενδιαφέρον. Το ANN έχει τα ακόλουθα χαρακτηριστικά. Πρώτον, το ANN είναι μία προσέγγιση που βασίζεται στα δεδομένα στην οποία υπάρχουν λιγότερες απαιτήσεις. Είναι ικανό να ανακαλύψει μία εσωτερική αναπαράσταση της σχέσης μεταξύ των δεδομένων και των μεταβλητών. Επομένως το μοντέλο ANN δεν απαιτεί υποθέσεις σχετικά με τη φύση της κατανομής των δεδομένων. Δεύτερον, το ANN έχει πολύ καλή μαθησιακή ικανότητα. Μετά από την προσαρμογή των δεδομένων, το ANN συλλαμβάνει αποτελεσματικά την φυσική κατασκευή και συμπεραίνει σωστά τον αφανή πληθυσμό. Επίσης είναι ικανό να προσεγγίσει μία μεγάλη ομάδα από συναρτήσεις με μεγάλη ακρίβεια. Τέλος, το ANN είναι ένας μηχανισμός μη γραμμικής αντιστοίχισης.

Το ANN είναι ένα παράλληλο σύστημα που αποτελείται από ιδιαίτερα διασυνδεδεμένα στοιχεία επεξεργασίας που βασίζονται σε νευροβιολογικά μοντέλα. Κατασκευάζει πληροφορίες μέσω αλληλεπιδράσεων ενός μεγάλου αριθμού στοιχείων, τους νευρώνες $(neurons)$. Οι κόμβοι των νευρωνικών δικτύων συνήθως χωρίζονται σε τρία στρώματα: τα εισαγόμενα, τα εξαγόμενα και τα τα κρυφά. Οι κόμβοι στα εσωτερικά στρώματα λαμβάνουν σήματα από εξωτερικές πηγές και οι κόμβοι στα εξωτερικά στρώματα παρέχουν το στόχο των σημάτων που εξάγονται. Η σχέση μεταξύ των στοιχείων που εξάγονται (y) και των στοιχείων που εισάγονται (x) σε ένα ANN μοντέλο είναι:

$$y = \alpha_0 + \sum_{i=1}^{b} \alpha_j g(\delta_{0_j} + \sum_{i=1}^{\alpha} \delta_{ij} x_i) + \varepsilon \tag{4.13}$$

όπου α_j $(j = 0, 1, 2, ..., b)$ και δ_{ij} $(i = 0, 1, 2, ..., a; j = 0, 1, 2, ..., b)$ είναι τα βάρη σύνδεσης του μοντέλου (*connection weights*), α είναι ο αριθμός των εισαγόμενων κόμβων, b είναι ο αριθμός των κρυφών κόμβων και ε είναι ο όρος σφάλματος. Η σύναρτηση μεταφοράς στο κρυφό στρώμα αναπαρίσταται από τη λογιστική συνάρτηση:

$$g(z) = \frac{1}{1 + exp(-z)} \tag{4.14}$$

4.4 Αποτελέσματα

Χωρίς βλάβη της γενικότητας, αυτή η μελέτη (Shao (2016)) θεωρεί ότι κάθε ποιοτικό χαρακτηριστικό αρχικά ακολουθεί την κανονική κατανομή με μέσο 0 και τυπική απόκλιση 1. Επειδή εξετάζονται 7 ποιοτικά χαρακτηριστικά για την πολυμεταβλητή κανονική διαδικασία, υπάρχουν $2^7 - 1$ πιθανές περιπτώσεις για τις μετατοπίσεις της διασποράς.

Συμβολίζονται με (1,0,...,0), (0,1,0,...,0),... και (1,1,...,1), όπου το 1 υποδηλώνει ένα ποιοτικό χαρακτηριστικό που είναι υπεύθυνο για το σφάλμα και το 0 υποδηλώνει ένα ποιοτικό χαρακτηριστικό που δεν είναι υπεύθυνο.

Για την κατασκευή ενός μη φυσικού διανύσματος διακύμανσης, θεωρούμε τις 2 συνηθισμένες περιπτώσεις αλλαγών διακύμανσης για επίδειξη (*demonstration*): (1,0,0,0,0,0,0) και (1,1,0,0,0,0,0). Επίσης θεωρούμε 4 διαφορετικές τιμές για το ϑ: ϑ=0.2, ϑ=0.4, ϑ=0.6, ϑ=0.8. Το μέγεθος του δείγματος είναι 10. Ο πίνακας διασποράς - συνδιασποράς για την περίπτωση της πολυμεταβλητής διεργασίας με 7 ποιοτικά χαρακτηριστικά περιγράφεται από τον

$$\mathbf{S} = \begin{bmatrix} S_{1,1} & S_{1,2} & \dots & S_{1,7} \\ S_{2,1} & S_{2,2} & \dots & S_{2,7} \\ \vdots & \vdots & \vdots & \vdots \\ S_{7,1} & S_{7,2} & \dots & S_{7,7} \end{bmatrix}$$

Σε αυτή την κατάσταση έχουμε 29 μεταβλητές στον πίνακα.

Στον Πίνακα 4.1 φαίνονται τα αποτελέσματα της επιλογής μεταβλητών μετά την εκτέλεση του $ANOVA$ στη φάση Ι. Όταν οι ποιοτικές μεταβλητές που είναι υπεύθυνες για το σφάλμα δεν είναι πολλές, η μελέτη θεωρεί ότι οι ποιοτικές μεταβλητές που είναι υπαίτιες είναι είτε η X_1 είτε οι X_1 και X_2.
Στον Πίνακα 4.2 φαίνεται η δομή των δεδομένων εκπαίδευσης και ελέγχου όπως προκύπτουν κατά την εφαρμογή του ANN. Στη συγκεκριμένη μελέτη περιλαμβάνονται 900 διανύσματα δεδομένων εκ των οποίων τα 300 πρώτα είναι εντός

Πίνακας 4.1: Αποτελέσματα επιλογής μεταβλητών με $ANOVA$.

θ	Variables selected
1.2	S_{11},S_{12},S_{13},S_{14},S_{15},S_{16},S_{17},S_{22},S_{33},S_{45}, S_{66}, SS (i.e., determinant of S)
1.4	S_{11},S_{12},S_{13},S_{14},S_{15},S_{16},S_{17},S_{22},S_{23},S_{24},S_{25}, S_{26},S_{27},S_{33},S_{44},S_{66},SS
1.6	S_{11},S_{12},S_{13},S_{14},S_{15},S_{16},S_{17},S_{22},S_{23},S_{24},S_{25}, S_{26},S_{27},S_{33},S_{36},S_{37},S_{44},S_{45},S_{66},SS
1.8	S_{11},S_{12},S_{13},S_{14},S_{15},S_{16},S_{17},S_{22},S_{23},S_{24},S_{25}, S_{26},S_{27},S_{33},S_{35},S_{36},S_{37},S_{44},S_{45},S_{66},SS

Πίνακας 4.2: Κατασκευή δεδομένων εκπαίδευσης και ελέγχου για το ANN.

Variables at fault	Output node (Y)	Number of observations
No	0	300
X_1	1	300
X_1 and X_2	2	300

ελέγχου, τα επόμενα 300 είναι εκτός ελέγχου και είναι υπεύθυνο το X_1 και τα τελευταία 300 είναι εκτός ελέγχου λόγω των X_1 και X_2.

Για ένα τυπικό στάδιο του μοντέλου ANN, υπάρχουν 29 εισαγόμενοι κόμβοι. Στο προτεινόμενο υβριδικό μοντέλο έχουμε 12, 17, 20 και 21 εισαγόμενους κόμβους για τα μοντέλα $ANOVA - ANN$ με τις περιπτώσεις ϑ=1.2, ϑ=1.4, ϑ=1.6 και ϑ=1.8, αντίστοιχα. Για όλα τα μοντέλα ο μόνος εξαγόμενος κόμβος είναι ο Y. Ο εξαγόμενος κόμβος υποδεικνύει τα αποτελέσματα της ταξινόμησης, όπου η τιμή 0 υποδηλώνει ότι η διεργασία είναι εντός ελέγχου και η τιμή 1 ότι είναι εκτός ελέγχου, όπου ευθύνεται η μεταβλητή X_1 και η τιμή 2 ότι είναι εκτός ελέγχου με υπεύθυνες τις μεταβλητές X_1, X_2.

Ο Πίνακας 4.3 περιέχει τα αποτελέσματα της προσομοίωσης για τις ακριβείς τιμές αναγνώρισης ($Accurate\ Identification\ Rates - AIR$) της τυπικής και προτεινόμενης προσέγγισης. Σύμφωνα με τον παρακάτω πίνακα η προτεινόμενη υβριδική προσέγγιση αποδίδει μόνο για το τυπικό ANN. Για παράδειγμα, για ϑ=1.2 και υπεύθυνες τις μεταβλητές X_1, X_2 η τιμή του AIR είναι 0.5333 και 0.6333 για την τυπική και την προτεινόμενη προσέγγιση αντίστοιχα, δηλαδή

έχει βελτίωση περίπου 19% .

Πίνακας 4.3: Αποτελέσματα προσομοίωσης για τις τιμές του AIR.

θ	Variable(s) at fault	Conventional Approach	Proposed hybrid approach
0.2	X_1	57.67%	72.00%
	X_1 and X_2	53.33%	63.33%
0.4	X_1	58.33%	66.00%
	X_1 and X_2	62.67%	63.67%
0.6	X_1	65.00%	64.00%
	X_1 and X_2	72.67%	74.33%
0.8	X_1	76.67%	74.67%
	X_1 and X_2	82.00%	87.67%

Ο Πίνακας 4.4 δείχνει τη γενική απόδοση, βάσει του AIR και της σχετικής τυπικής απόκλισης, για την προτεινόμενη υβριδικη και τη συνηθισμένη προσέγγιση. Η ποιοτική μεταβλητή X_1 σε υπαιτιότητα μπορεί με ακρίβεια να αναγνωριστεί με 64.42% πιθανότητα επιτυχίας με τη χρήση της συνηθισμένης ANN προσέγγισης. Η σχετική τυπική απόκλιση είναι 8.81%. Επιπρόσθετα υποδηλώνει ότι η ποιοτική μεταβλητή X_1 μπορεί να αναγνωριστεί με ακρίβεια με πιθανότητα 69.17% κάνοντας χρήση του προτεινόμενου υβριδικού μοντέλου και η σχετική τυπική απόκλιση είναι 5.00%.

Συνεπώς, με τη χρήση της προτεινόμενης προσέγγισης αντί για την συνηθισμένη μπορούμε να πετύχουμε 7.37% βελτίωση του AIR και μείωση ή βελτίωση της σχετικής τυπικής απόκλισης κατά 43.62%.

Σχετικά με τα σφάλματα:

Όταν ένα μοντέλο κατασκευάζεται από δεδομένα κανονικής διεργασίας με τη χρήση της μεθόδου PCA ή ICA, μπορεί να χρησιμοποιηθεί για την ανίχνευση και την αναγνώριση μη συνηθισμένων καταστάσεων σε μία διεργασία, όπως είναι τα σφάλματα της διεργασίας και των αισθητήρων ($sensors$). Τα στατιστικά παρακολούθησης χρησιμοποιούνται για την ανίχνευση σφαλμάτων και τα διαγράμματα συνεισφοράς συνήθως χρησιμοποιούνται για την αναγνώριση τους.

Πίνακας 4.4: Απόδοση για τις δύο προσεγγίσεις.

	AIR	Standard deviation
Conventional approach		
X_1 at fault	64.42%	8.81%
X_1 and X_2 at fault	67.67%	12.40%
Proposed hybrid approach		
X_1 at fault	69.17%	5.00%
X_1 and X_2 at fault	72.25%	11.48%

Τέλος γίνεται μία αναφορά στην ενσωμάτωση της ανάλυσης ανεξάρτητων συνιστωσών (ICA) και των μηχανών διανυσμάτων υποστήριξης (SVM) για την παρακολούθηση πολυμεταβλητών διεργασιών. Για τη δημιουργία ενός επιτυχημένου ανιχνευτή σφαλμάτων που βασίζεται στο SVM, το πρώτο στάδιο είναι η εξαγωγή χαρακτηριστικών. Στις πραγματικές βιομηχανικές διαδικασίες, οι μεταβλητές σπάνια ακολουθούν την $Gaussian$ κατανομή. Έτσι η συγκεκριμένη μελέτη προτείνει την εφαρμογή της ICA για να εξαχθούν οι κρυφές πληροφορίες μίας $non-Gaussian$ διεργασίας πριν τη χρήση των SVM. Η ICA αρχικά ανακαλύφθηκε για εφαρμογές επεξεργασίας σήματος, συμπεριλαμβανομένου και της επεξεργασίας σήματος ομιλίας, επικοινωνίας, επεξεργασίας ιατρικών εικόνων κ.α. Η ICA θεωρείται σαν μία επέκταση της PCA αν και το αντικείμενο των δύο αλγορίθμων είναι κάπως διαφορετικό. Η PCA μπορεί να επιβάλλει την ανεξαρτησία σε στατιστικά δεύτερης τάξης, όπως η συνδιακύμανση και έτσι το αντικείμενό της είναι να αποσυσχετίζει μεταβλητές. Αντιθέτως η ICA επιβάλλει μία στατιστική ανεξαρτησία στις ατομικές συνιστώσες θεωρώντας στατιστικά υψηλής τάξης. Η ICA παρέχει περισσότερες πληροφορίες από την PCA.

Οι Kano et al. (2003) σύγκριναν τα αποτελέσματα της παρακολούθησης εφαρμόζοντας SPC διαγράμματα στις συνιστώσες της PCA και της ICA αντίστοιχα. Τα αποτελέσματα έδειξαν ότι είναι πιο αποτελεσματική η παρακολούθηση ICA συνιστωσών από την παρακολούθηση PCA συνιστωσών όταν η συμεριφορά των μεταβλητών της διεργασίας ακολουθούν $non-Gaussian$ κατανομή. Αν και τα αποτελέσματα έδειξαν ότι η τεχνική ICA μπορεί να παρακολουθεί $non-Gaussian$ διεργασίες, μπορεί να παράγει λανθασμένες ειδοποιήσεις, όταν οι συνιστώσες παρακολουθούνται μεμονωμένα από διαγράμματα SPC. Έτσι οι Lee et al. (2004) ανακάλυψαν τρία στατιστικά παρακολούθησης που βασίζονται στην ICA για να παρακολουθεί τη διεργασία. Επιπλέον, οι Lee et al. (2006) πρότειναν ένα τροποποιημένο ICA για να ξεπεράσουν τα μειονεκτήματα του αρχικού ICA αλγόριθμου, όπως ο προκαθορισμός και των αριθμών

των εξαγόμενων ανεξάρτητων συνιστωσών και της σωστής σειράς των ανεξάρτητων συνιστωσών. Οι Yoo et al. (2004) ανακάλυψαν ένα πολλαπλό σχήμα παρακολούθησης βασιζόμενο στην ICA για την παρακολούθηση του συνόλου παραγωγής. Οι Ge and Song (2007) πρότειναν μία $PCA - ICA$ μέθοδο που εξάγει $Gaussian$ και $non - Gaussian$ πληροφορίες για ανίχνευση σφαλμάτων.

Τα όρια ελέγχου για ένα στατιστικό παρακολούθησης βασιζόμενο στην ICA δεν μπορούν να καθοριστούν από κάποια συγκεκριμένη κατανομή. Παραδοσιακά η εκτίμηση πυκνότητας πυρήνα ($Kernel\ Density\ Estimation - KDE$) χρησιμοποιείται για να καθορίσει τα όρια ελέγχου για τα στατιστικά παρακολούθησης που βασίζονται στην ICA. Η KDE έχει δύο περιορισμούς, την απαίτηση μεγάλων ομάδων δεδομένων και υψηλή ευαισθησία στην επιλογή παραμέτρων εξομάλυνσης. Πρόσθετα δεν ταιριάζει καλά σε δεδομένα που έχουν αυτοσυσχέτιση. Ακόμα και αν η χρήση της δυναμικής ICA ($Dynamic\ ICA$) μπορεί να εξαλείψει την αυτοσυσχέτιση των δεδομένων, ο πίνακας προεπεξεργασμένων δεδομένων πρέπει να επεκταθεί για να περιέχει τις μεταβλητές που εισέρχονται με χρονική καθυστέρηση, το οποίο θα αυξήσει κατά πολύ την πολυπλοκότητα των υπολογισμών της ICA και θα απαιτεί περισσότερες ανεξάρτητες συνιστώσες να εξαχθούν από την ανάλυση.

Στη συγκεκριμένη έρευνα Hsu et al. (2010) προτείνεται ένας νέος έξυπνος ανιχνευτής σφαλμάτων που ενσωματώνει δύο ιστορικές τεχνικές που έχουν σαν οδηγό τα δεδομένα, την ICA και το SVM. Το SVM είναι μία μέθοδος μάθησης που δεν απαιτεί υποθέσεις για την κατασκευή των δεδομένων και χρησιμοποιείται ευρέως για προβλήματα ταξινόμησης. Άλλες έρευνες έχουν δείξει πλεονεκτήματα στην ενσωμάτωση της ανάλυσης των συνιστωσών στο SVM σε πολλές εφαρμογές. Ωστόσο οι έρευνες αυτές χρησιμοποιούν εξαγόμενες συνιστώσες σαν εισαγωγή στο SVM. Οπότε σε αύτη την έρευνα προτείνεται ο συνδυασμός των συνιστωσών σε ένα στατιστικό σαν χαρακτηριστικό εισαγωγής στο SVM. Η βασική ιδέα είναι η εξής: Αρχικά χρησιμοποιείται η ICA για να μειώσει τις διαστάσεις και να εξάγει ανεξάρτητες συνιστώσες. Έπειτα οι ανεξάρτητες συνιστώσες χρησιμοποιούνται για να υπολογιστεί το συστηματικό στατιστικό. Για να ληφθεί και η αυτοσυσχέτιση υπόψη, τα δεδομένα που εισάγονται στο SVM είναι:

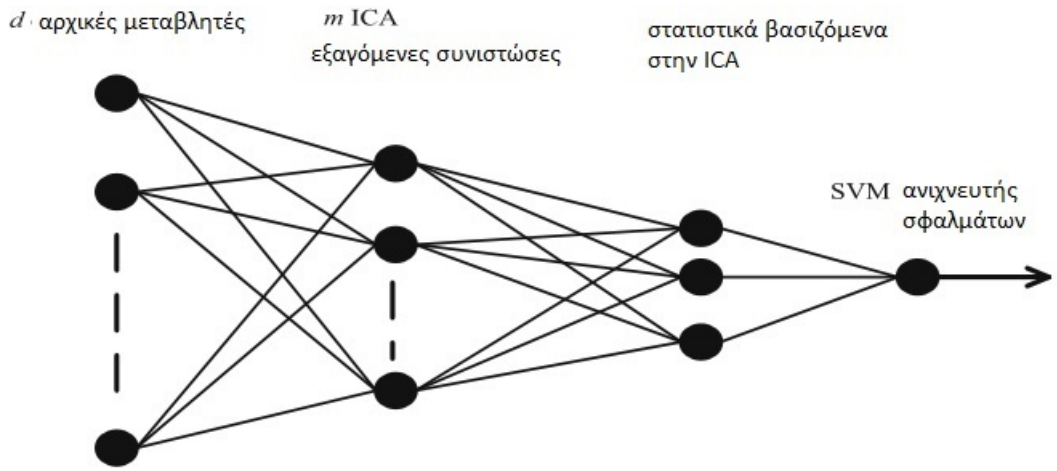
1. το συστηματικό στατιστικό του παρόντος χρόνου,

2. το συστηματικό στατιστικό με μία μικρή καθυστέρηση,

3. η διαφορά μεταξύ δύο επιτυχημένων συστηματικών στατιστικών.

Τα αποτελέσματα δείχνουν ότι το προτεινόμενο μοντέλο κατέχει ανώτερη ανίχνευση σφαλμάτων σε σχέση με άλλα μοντέλα όπως τα PCA, ICA, $MICA$ ($Modified\ ICA$), $ICA - PCA$ και $PCA - SVM$.

Στο Σχήμα 4.2 φαίνεται η αρχιτεκτονική της προτεινόμενης μεθοδολογίας. Αρχικά τα εξαγόμενα χαρακτηριστικά που βασίζονται στην ICA χρησιμοποιούνται για να προβάλλουν την ομάδα δεδομένων υψηλής διάστασης σε ένα σετ μικρότερης διάστασης. Οι εξαγόμενες ICs χρησιμοποιούνται στη συνέχεια το στατιστικό συστηματικό κομμάτι. Για να ληφθεί υπόψη η αυτοσυσχέτιση, η καθυστέρη του χρόνου και η διαφορά του χρόνου των συστηματικών συστατικών θεωρούνται επίσης σαν διανύσματα εισαγωγής για το $ICA - SVM$. Η ανάπτυξη του ανιχνευτή σφαλμάτων $ICA - SVM$ περιέχει δύο φάσεις, την $off - line$ εκπαίδευση και τον $on - line$ έλεγχο.



Σχήμα 4.2: Η αρχιτεκτονική του $ICA - SVM$ ανιχνευτή σφαλμάτων.

5 Διαγράμματα ελέγχου $SVM - MSPC$

5.1 Εισαγωγή

Τα παραδοσιακά πολυμεταβλητά διαγράμματα ελέγχου θεωρούν ότι οι μετρήσεις της παραγωγικής διεργασίας ακολουθούν μία πολυμεταβλητή κανονική κατανομή. Αυτή η υπόθεση μπορεί να μην ισχύει ή να είναι πολύ δύσκολο να επαληθευτεί, επειδή πρακτικά όλες οι μετρήσεις από τις παραγωγικές διεργασίες δεν είναι κανονικά κατανεμημένες.

Η λειτουργία των τεχνικών του SPC είναι να αναγνωρίζουν τις αλλαγές στις διεργασίες παραγωγής, ώστε οι υπεύθυνοι για τη λήψη αποφάσεων να μπορούν να λάβουν διορθωτικά μέτρα πριν αλλοιωθεί η ποιότητα. Στην πραγματικότητα τα δεδομένα είναι συχνά πολυμεταβλητά και συσχετισμένα. Έτσι σε αυτές τις διεργασίες εφαρμόζονται οι τεχνικές του $MSPC$. Είναι σύνηθες να παρακολουθούνται ταυτόχρονα διάφορες μετρήσεις μιας διεργασίας. Ένα σημαντικό εργαλείο του $MSPC$ είναι τα πολυμεταβλητά διαγράμματα ελέγχου (*multivariate control charts*) που χρησιμοποιούνται για να ανιχνεύουν τις μετατοπίσεις και να διατηρούν τη διεργασία σε μία εντός ελέγχου κατάσταση.

Κάποια πολυμεταβλητά διαγράμματα ελέγχου όπως το T^2, το $MCUSUM$ και το $MEWMA$ που έχουν αναφερθεί σε προηγούμενη ενότητα συνήθως θεωρούν ότι οι μετρήσεις ακολουθούν μία πολυμεταβλητή κανονική κατανομή και σε πολλές περιπτώσεις αυτή η υπόθεση είναι δύσκολο να επαληθευτεί. Για την παρακολούθηση της ποιότητας μίας διεργασίας με μη-κανονικά δεδομένα, έχουν προταθεί μη παραμετρικά πολυμεταβλητά διαγράμματα ελέγχου, όπως το $Antirank-based\ CUSUM$ ($ACUSUM$) (Qiu and Hawkins, 2001, 2003), το $Multivariate\ Sign\ EWMA$ ($MSEWMA$) (Zou and Tsung, 2011) και άλλα. Αυτά τα διαγράμματα ελέγχου απαιτούν λιγότερο υπολογιστικό χρόνο και είναι πιο ευαίσθητα στις αλλαγές. Όμως είναι δύσκολο να κατασκευαστούν γιατί ο καθορισμός των παραμέτρων βάρους εξαρτώνται από την απόκλιση της κατανομής των πραγματικών μετρήσεων από την πολυ-κανονική ($multi-normal$) κατανομή. Πρόσθετα το μη παραμετρικό ($rank-based$) διάγραμμα ελέγχου είναι συνήθως μη αποτελεσματικό για μεγάλες μετατοπίσεις.

Κάποιες μελέτες έχουν ανακαλύψει ένα διάγραμμα ελέγχου βασίζεται στη μηχανική μάθηση (*machine learning*) ή στην εξόρυξη δεδομένων (*data mining*). Οι Sun and Tsung (2003) προτείνουν τα K διαγράμματα που βασίζονται στον αλγόριθμο περιγραφής δεδομένων διανυσμάτων υποστήριξης ($Support\ Vector\ Data\ Description$-$SVDD$) για τα οποία θα γίνει αναφορά παρακάτω. Οι Kumar et al. (2006), Camci et al. (2008), Cheng and Cheng (2008) προσπάθησαν να βελτιώσουν το σχεδιασμό των K διαγραμμάτων. Οι Ning and Tsung (2013) πρότειναν ένα συστηματικό σχεδιασμό των K διαγραμμάτων και έφτιαξαν μία

ολοκληρωμένη ανάλυση του σχεδιασμού του K διαγράμματος. Πρόσθετα οι Sukchotrat et al. (2009) παρουσίασαν ένα K^2 διάγραμμα που βασίζεται στον αλγόριθμο περιγραφής δεδομένων των k κοντινότερων γειτόνων ($kNNDD$). Παρόλα αυτά, τα περισσότερα από αυτά τα διαγράμματα ελέγχου είναι δύσκολο να σταθεροποιηθούν, γιατί ο καθορισμός των παραμέτρων είναι απαραίτητος και ο τρόπος καθορισμού αυτών των παραμέτρων σπάνια είναι γνωστός.

Παρακάτω προτείνεται μία καινούρια μεθοδολογία πολυμεταβλητού SPC για την παρακολούθηση της διεργασίας με μη κανονικά δεδομένα. Αυτή η μεθοδολογία βασίζεται στην ολοκλήρωση (*integrating*) της μεθόδου ταξινόμησης μίας κλάσης και στην προσαρμοσμένη τεχνική (*adaptive technique*). Η προσαρμοσμένη τεχνική χρησιμοποιείται για να βελτιώσει την ευαισθησία στις μικρές αλλαγές στην ταξινόμηση μίας τάξης στο στατιστικό έλεγχο διεργασιών. Επιπρόσθετα αυτός ο σχεδιασμός παρέχει έναν εύκολο τρόπο να διανέμεται η τιμή του σφάλματος τύπου I έτσι ώστε να είναι πιο εύκολο να εφαρμοστεί.

5.2 Το διάγραμμα ελέγχου που βασίζεται στην $SVDD$

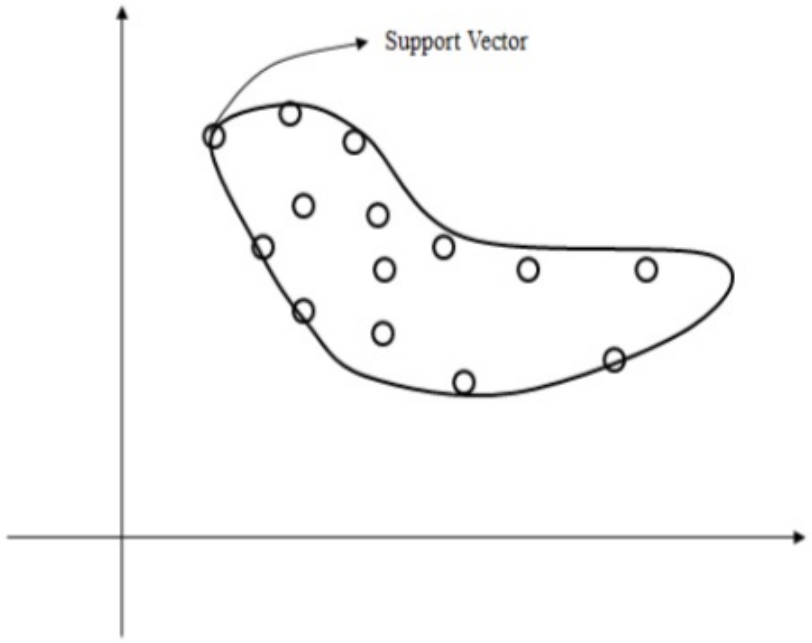
Η $SVDD$ είναι ένας αλγόριθμος ταξινόμησης μίας κλάσης, ο οποίος χρησιμοποιείται για τον SPC. Ειναι μία μίξη του SVM και της μεθόδου περιγραφής δεδομένων για τη λύση προβλημάτων ταξινόμησης μίας κλάσης.

Περιγραφή της $SVDD$:

Συνήθως το SVM χρησιμοποιείται για να λύνει δυαδικά προβλήματα ταξινόμησης. Στον έλεγχο διεργασιών όμως δεν έχουμε δεδομένα δύο κλάσεων αλλά εντός ελέγχου δεδομένα. Έτσι η μέθοδος SVM δεν μπορεί να εφαρμοστεί για την παρακολούθηση της διεργασίας. Οι Tax and Duin (1999) πρότειναν πρώτοι τη μέθοδο $SVDD$ που προέρχεται από το SVM και τη χρησιμοποίησαν για να λύσουν προβλήματα ταξινόμησης μίας κλάσης και στη συνέχεια χρησιμοποιήθηκε για την παρακολούθηση διεργασιών.

Η βασική ιδέα της $SVDD$ είναι να βρει μία υπερσφαίρα λύνοντας το πρόβλημα βελτιστοποίησης, το οποίο ελαχιστοποιεί τον όγκο της υπερσφαίρας και μεγιστοποιεί το πλήθος των δεδομένων εκπαίδευσης που περιέχονται στην υπερσφαίρα.

Έστω ότι X_i=$[x_1, x_2, ..., x_p]^T$ με $i = 1, 2, ..., N$ η ομάδα εκπαίδευσης. Σημειώνεται ότι υπάρχει μόνο μία κλάση. Δίνεται το κέντρο της υπερσφαίρας, έστω O, και η ακτίνα της, έστω R. Το μοντέλο βελτιστοποίησης φαίνεται παρακάτω:



Σχήμα 5.1: Η βασική ιδέα της $SVDD$

$$min R^2 + C\sum_{i=1}^{N}\xi_i \tag{5.1}$$

$$\text{δεδομένου ότι } (X_i - O)^T(X_i - O) \leq R^2 + \xi_i \tag{5.2}$$

$$\xi_i \geq 0, \forall i, \tag{5.3}$$

όπου το ξ_i είναι η μεταβλητή απόκλισης ($slack\ variable$) που αντιστοιχεί στους περιορισμούς. Το $C > 0$ είναι η ανταλλαγή ($trade - off$) μεταξύ του όγκου της υπερσφαίρας και των σημείων των δεδομένων εκπαίδευσης που περιέχονται στην υπερσφαίρα. Αν το C είναι μεγάλο, το αντίστοιχο σύνορο θα είναι ευρύ και θα περιέχονται πολλά σημεία στην υπερσφαίρα και αν το C είναι μικρό θα περιέχονται λιγότερα σημεία.

Το μοντέλο αυτό μπορεί να λυθεί με τη χρήση της $Langrangian$ συνάρτησης:

$$L(R, O, \pi_i, \tau_i, \xi_i) = R^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\pi_i[R^2 + \xi_i - (X_i - O)^T(X_i - O)] - \sum_{i=1}^{N}\tau_i\xi_i, \tag{5.4}$$

όπου π_i και τ_i είναι οι πολλαπλασιαστές $Lagrange$ και είναι μεγαλύτεροι ή ίσοι με 0. Οπότε οι καινούριοι περιορισμοί είναι:

$$2R - 2R\sum_{i=1}^{N}\pi_i = 0 \tag{5.5}$$

$$-2\sum_{i=1}^{N}\pi_i(X_i - O) = 0 \tag{5.6}$$

$$C - \pi_i - \xi_i = 0. \tag{5.7}$$

Έτσι το μοντέλο παίρνει την απλοποιημένη μορφή:

$$max\sum_{i=1}^{N}\pi_i X_i^T X_i - \sum_{i,j=1}^{N}\pi_i\pi_j X_i^T X_j \tag{5.8}$$

$$\text{δεδομένου ότι } \sum_{i=1}^{N}\pi_i = 1 \tag{5.9}$$

$$0 \leq \pi_i \leq C, \forall i. \tag{5.10}$$

Στη συνέχεια μπορούμε να καθορίσουμε αν το σημείο ελέγχου, Z, ανήκει στην ομάδα (δηλαδή εντός ελέγχου δεδομένα), υπολογίζοντας την απόσταση μεταξύ του σημείου ελέγχου, Z, και του κέντρου, O. Αν η απόσταση ($\|Z-O\|^2$) είναι μεγαλύτερη από την ακτίνα R, αυτό αναγνωρίζεται σα μη-φυσικό σημείο και μπορεί να διατυπωθεί ως εξής:

$$\|Z-O\|^2 = (Z-\sum_{i=1}^{N}\pi_i X_i)^T(Z-\sum_{i=1}^{N}\pi_i X_i) = Z^T Z - 2\sum_{i=1}^{N}\pi_i X_i^T Z + \sum_{i,j=1}^{N}\pi_i\pi_j X_i^T X_j. \tag{5.11}$$

Στη συνέχεια μπορούμε να αντικαταστήσουμε το εσωτερικό γινόμενο με συναρτήσεις πυρήνα, ώστε κάποιος να έχει ένα πιο προσαρμοσμένο σύνορο του $SVDD$ σε σύγκριση με το $SVDD$ χωρίς τη συνάρτηση πυρήνα. Μία γνωστή συνάρτηση πυρήνα είναι η $Gaussian\ RBF$ ($Gaussian\ Radical\ Basis\ Function$):

$$K(X_i, X_j) = exp(-\frac{\|X_i - X_j\|^2}{g^2}), \tag{5.12}$$

όπου $g > 0$ είναι το πλάτος του παραθύρου ($Gaussian\ kernel\ that\ controls\ the\ complexity\ of\ the\ SVDD\ boundary$). Αντικαθιστούμε με τη συνάρτηση

Gaussian RBF:

$$max \sum_{i=1}^{N} \pi_i K(X_i, X_j) - \sum_{i,j=1}^{N} \pi_i \pi_j K(X_i, X_j) \tag{5.13}$$

$$\text{δεδομένου ότι } \sum_{i=1}^{N} \pi_i = 1 \tag{5.14}$$

$$0 \leq \pi_i \leq C, i = 1, ..., N. \tag{5.15}$$

Και η απόσταση του πυρήνα ($D^2(Z)$) μεταξύ του σημείου ελέγχου και του κέντρου είναι:

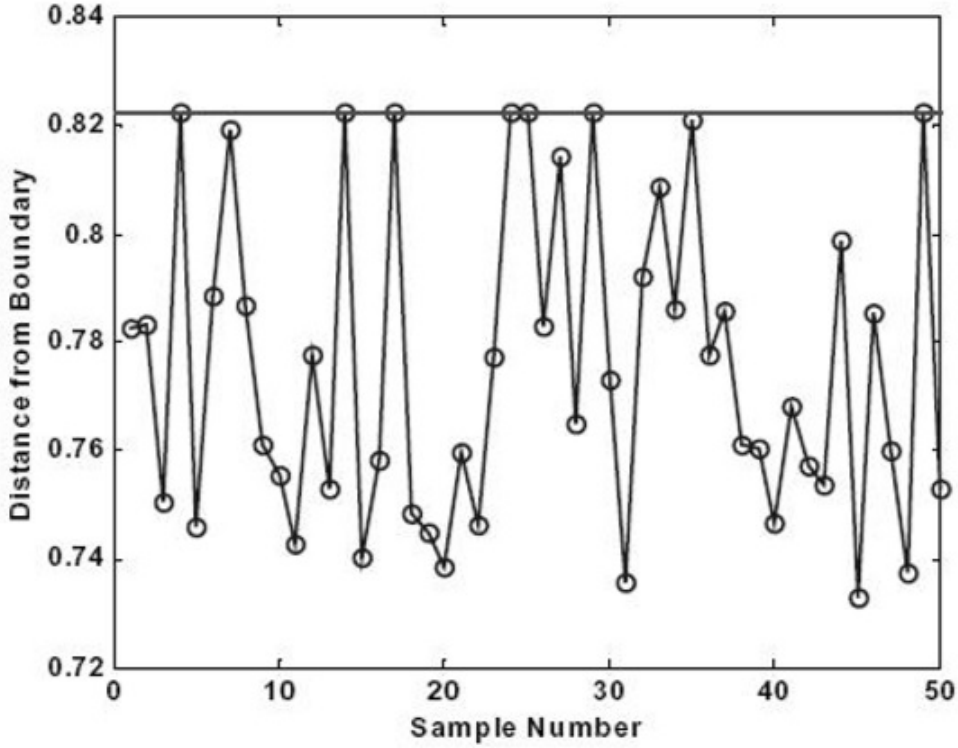
$$D^2(Z) = K(Z, Z) - 2 \sum_{i=1}^{N} \pi_i K(Z, X_i) + \sum_{i,j=1}^{N} \pi_i \pi_j K(X_i, X_j). \tag{5.16}$$

Τέλος, η $D^2(Z)$ μπορεί να χρησιμοποιηθεί για να δείξει ότι τα δεδομένα ελέγχου είναι σε μία συγκεκριμένη κατηγορία.

5.3 Διαγράμματα K (K-charts)

Βάσει του $SVDD$, οι Sun and Tsung (2003) πρότειναν το K-διάγραμμα, το οποίο βασίζεται σε έναν $SVDD$ αλγόριθμο (Σχήμα 5.2). Το K διάγραμμα βασίζεται στις μηχανές διανυσμάτων υποστήριξης που ανακαλύφθηκαν για την παρακολούθηση πολυμεταβλητών διεργασιών. Το διάγραμμα ελέγχου αναζητά τα διανύσματα υποστήριξης με το SVM και εξασφαλίζει την απόσταση πυρήνα (*kernel distance*) για να σταθεροποιηθεί.

Οι Sun and Tsung (2003), στο παράδειγμα τους υποθέτουν ότι $C > 1$. Πιο συγκεκριμένα, η γενική μορφή του $SVDD$, εξίσωση (5.13), έχει δύο περιορισμούς. Με τον περιορισμό $\sum_{i=1}^{N} \pi_i = 1$ και $\pi_i \geq 0$, $i = 1, ..., N$, έχουμε ότι $0 \leq \pi_i \leq 1$, $i = 1, ..., N$. Και αν $C > 1$, με $0 \leq \pi_i \leq C$, έχουμε ότι $0 \leq \pi_i \leq 1$, $i = 1, ..., N$. Έτσι η ανισότητα του περιορισμού μπορεί να χαλαρώσει, και η



Σχήμα 5.2: Το K διάγραμμα.

εξίσωση (5.13) να απλουστευτεί σε:

$$max \sum_{i=1}^{N} \pi_i K(X_i, X_j) - \sum_{i,j=1}^{N} \pi_i \pi_j K(X_i, X_j) \tag{5.17}$$

$$\text{δεδομένου ότι } \sum_{i=1}^{N} \pi_i = 1 \tag{5.18}$$

$$\pi_i \geq 0, i = 1, ..., N, \tag{5.19}$$

το οποίο χρησιμοποιείται από τους Sun and Tsung (2003). Ο συντελεστής ποινής C δεν εμφανίζεται εδώ. Σημειώνεται επίσης ότι η απόσταση πυρήνα για ένα σημείο ελέγχου, έστω Z, καθορίζεται με τον ίδιο τρόπο, όπως στην εξίσωση (5.16). Όσο πιο μακριά είναι ένα σημείο από το σύνολο εκπαίδευσης, τόσο μεγαλύτερη είναι αυτή η απόσταση. Δεδομένου ενός σημείου-κατώφλι, h, στο $\{D^2(X_i), i = 1, ..., N\}$, αν ένα σημείο Z, ικανοποιεί την ανισότητα $D^2(Z) > h$, θεωρείται ότι είναι εκτός ελέγχου. Αυτή η τιμή-κατώφλι, h, θεωρείται σαν το όριο ελέγχου για το $\{D^2(X_i), i = 1, ..., N\}$.

Πλεονεκτήματα και μειονεκτήματα των K-διαγραμμάτων.

Στην πραγματικότητα το K-διάγραμμα έχει μερικά σημαντικά πλεονεκτήματα. Το πρώτο απ'αυτά είναι ότι το K-διάγραμμα δεν απαιτεί καμία υπόθεση σχετικά με την κατανομή που ακολουθούν τα δεδομένα, το οποίο είναι θεμελιώδες για τα υπόλοιπα διαγράμματα. Το δεύτερο πλεονέκτημα των K-διαγραμμάτων είναι η ικανότητά τους να αντιμετωπίζουν μεγάλο πλήθος μεταβλητών χωρίς να αλλοιώνεται η αποτελεσματικότητά τους. Σύμφωνα με τον Montgomery (2013), όταν η διάσταση των μεταβλητών αυξάνεται, τα περισσότερα SPC χάνουν την αποτελεσματικότητά τους σχετικά με την ανίχνευση μετατοπίσεων. Η ικανότητα του K-διαγράμματος να χειρίζεται ένα μεγάλο αριθμό μεταβλητών οφείλεται στις μεθόδους πυρήνα.

Ανεξάρτητα όμως από τα πολλά πλεονεκτήματα, τα K-διαγράμματα έχουν και μερικά μειονεκτήματα. Το πρώτο σχετίζεται με τον καθορισμό μίας εντός ελέγχου διεργασίας. Σύμφωνα με μελέτες, δεν είναι εύκολο να κατασκευαστεί μία εντός ελέγχου κατάσταση ή να διαπιστωθεί ότι η διεργασία είναι εντός ελέγχου. Το πρόβλημα σχετίζεται με τον καθορισμό της βέλτισης κλάσης-στόχου, η οποία αποτελεί τη βάση της εντός ελέγχου κατάστασης για το K-διάγραμμα. Αυτή η κλάση-στόχος μπορεί να περιέχει έκτροπες τιμές αν οι παράμετροι του $SVDD$ ταξινομητή δεν έχουν επιλεχθεί με προσοχή. Για αυτόν το λόγο, η βέλτιστη κλάση-στόχος απαιτεί περαιτέρω βελτιστοποίηση για να βεβαιωθεί ότι αντιπροσωπεύει αποτελεσματικά την εντός ελέγχου διεργασία.

Ένα άλλο θέμα σχετικά με το K-διάγραμμα αφορά κυρίως τα όρια ελέγχου. Το γεγονός ότι τα όρια ελέγχου βασίζονται στα διανύσματα υποστήριξης, αυξάνει την πιθανότητα ένα διάνυσμα υποστήριξης να είναι έκτροπη τιμή. Αυτό μπορεί να εξηγηθεί από το γενονός ότι το κέντρο του πυρήνα είναι ένας γραμμικός συνδυασμός των διανυσμάτων υποστήριξης. Άλλωστε, ο αριθμός των διανυσμάτων υποστήριξης που χρειάζεται για την κατασκευή των ορίων ελέγχου παρουσιάζει ένα άλλο θέμα που απαιτεί έρευνα.

Σχετικά με τον $SVDD$ ταξινομητή, απαιτούνται 50 με 100 παρατηρήσεις για να έχουμε μία καλή περιγραφή δεδομένων. Λαμβάνοντας υπόψη ότι αυτός ο αριθμός των παρατηρήσεων χρησιμοποιείται για να κατασκευαστεί η εντός ελέγχου κατάσταση, μπορεί να προκληθεί πρόβλημα μεροληψίας στην επιλογή των παρατηρήσεων, σε σύγκριση με τα παραδοσιακά διαγράμματα ελέγχου που απαιτούν 15 με 25 δείγματα για να κατασκευάσουν μία εντός ελέγχου διεργασία.

5.3.1 Το βελτιωμένο K-διάγραμμα

Σημαντικό ρόλο στην αποτελεσματικότητα των K-διαγραμμάτων παίζει ο καθορισμός των παραμέτρων g και C. Οι Ning and Tsung (2013) μελέτησαν το πως γίνεται να υπολογιστούν αυτές οι δύο παράμετροι. Η ιδέα τους είναι να βρεθούν δύο παράμετροι που έχουν το ελάχιστο ολικό σφάλμα, γ. Το γ ισούται με το σταθμισμένο άθροισμα ($weighted\ sum$) του σφάλματος τύπου-I και του σφάλματος τύπου-II, το οποίο είναι:

$$\gamma = (1 - v) \times \frac{\#SV}{N} + v \times f_{o+}, \tag{5.20}$$

όπου το v δείχνει τη σημαντικότητα μεταξύ των σφαλμάτων τύπου I και II. Τα σφάλματα τύπου I υπολογίζονται από το ποσοστό του αριθμού των $Supper$ διανυσμάτων (SV), $\#SV$, και τον αριθμό των δεδομένων εκπαίδευσης. Τα σφάλματα τύπου II υπολογίζονται από το ποσοστό του αριθμού των τεχνητών έκτροπων τιμών ($artificial\ outlier$) που αναγνωρίζονται σαν εντός ελέγχου δεδομένα. Λόγω του ότι είναι δύσκολο να υπολογιστεί η σημαντικότητα μεταξύ των σφαλμάτων τύπου I και II, οι χρήστες θεωρούν περιττό να καθορίσουν το βάρος (v). Επομένως είναι δύσκολο να χρησιμοποιηθούν τα διαγράμματα K.

Ένα άλλο πρόβλημα στη ανάλυση της φάσης I είναι ο τρόπος με τον οποίο θα αναπαραστήσουμε τις πιθανές ακραίες τιμές. Στο βελτιωμένο K-διάγραμμα χρησιμοποιείται η μέθοδος $leave-one-out\ cross-validation$ για να εντοπίσει τις ακραίες τιμές. Η μέθοδος $cross-validation$ χρησιμοποιείται συχνά στον έλεγχο υποθέσεων. Διαχωρίζοντας τις δεδομένες ομάδες σε δύο μέρη, το ένα να είναι το κομμάτι εκπαίδευσης και το άλλο το κομμάτι ελέγχου, έχουμε ένα τρόπο να βρούμε τα περίεργα σημεία.

Έστω ένα σύνολο εκπαίδευσης, $X_1, ..., X_n$, και ένα σημείο, X_i, $\forall i \in 1, ..., n$, επιλέγεται σαν σημείο ελέγχου. Τότε, όλα τα άλλα σημεία χρησιμοποιούνται για να εκτιμήσουν την απόσταση πυρήνα. Έστω

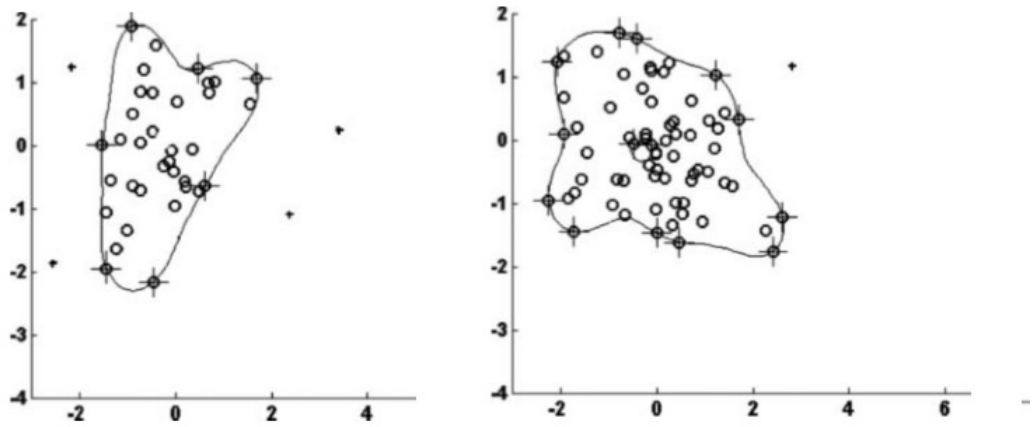
$$r(i) = \frac{D^2(X_i)}{max_{k=1,...,n,k\neq i}[D^2(X_k)]} \tag{5.21}$$

και

$$r_{max} = max_{i=1,...,n} r(i). \tag{5.22}$$

Αν το σημείο που αντιστοιχεί στο r_{max} βρίσκεται σε εκτός ελέγχου κατάσταση, το διαγράφουμε. Στη συνέχεια εξετάζουμε το καινούριο σύνολο, το οποίο έχει $n - 1$ σημεία, και επαναλαμβάνουμε τα παραπάνω βήματα, μέχρι να μην διαγράφονται πλέον σημεία. Με αυτόν τον τρόπο μπορούμε να βρούμε τα πιο μακρινά σημεία και να ελένξουμε αν η συμπεριφορά τους οφείλεται σε προσδιορισμένα αίτια.

Παρακάτω δίνεται ένα παράδειγμα εφαρμογής αυτής της μεθόδου, όπου έχουμε διδιάστατα δεδομένα από το $N(\mathbf{0}, \mathbf{I}_2)$ και τα οποία χρησιμοποιούνται για να απεικονίσουμε την απόδοση της. Το μέγεθος του εκπαιδευόμενου δείγματος έχει επιλεχθεί να είναι $n = 40$ και 70. Το Σχήμα 5.3 δείχνουν το αποτέλεσμα του ορίου ελέγχου μετά τη διαγραφή των ακραίων σημείων με τη χρήση της $leave - one - out\ cross - validation$ μεθόδου. Αυτά τα σχήματα δείχνουν ότι αυτή η μέθοδος είναι εφαρμόσιμη στην εύρεση πιθανών ακραίων τιμών.



Σχήμα 5.3: Τα αποτελέσματα της μεθόδου $leave-one-out\ cross-validation$ στα όρια ελέγχου για $n = 40$ (αριστερά) και $n = 70$ (δεξιά).

Ο συντελεστής ποινής, C, ελέγχει τον αριθμό των σημείων που πέφτουν εκτός του ορίου. Οι Schoelkopf et al. (2001) τροποποίησαν το πρόβλημα (5.13) αντικαθιστώντας το συντελεστή ποινής, C, με $1/vn$ και ονόμασαν την τροποποιημένη μορφή v-διάνυσμα υποστήριξης ($v - support\ vector$). Πιο συγκεκριμένα, έχει την ακόλουθη μορφή:

$$min R^2 + \frac{1}{vn}\sum_{i=1}^{n}\xi_i, \tag{5.23}$$

$$\text{δεδομένου ότι } \|\mathbf{O} - \mathbf{X}_i\|^2 \leq R^2 + \xi_i^2 \tag{5.24}$$

$$\xi \geq 0, i = 1, ..., n, \tag{5.25}$$

όπου $\mathbf{O}$ είναι το κέντρο, R είναι η ακτίνα, το ξ_i είναι οι χαλαρές μεταβλητές και $v \in [0, 1]$. Εδώ το v είναι ανεξάρτητη παράμετρος. Είναι ένα άνω όριο για τα σημεία που ταξινομούνται σαν ειδικά σημεία είτε σαν ύποπτα σημεία.

Συγκρίνοντας με το C, το v είναι πιο κατανοητό, διότι είναι είναι αντίστοιχο του σφάλματος τύπου Ι. Το αντίστοιχο δυϊκό πρόβλημα του $v - SVM$ με την εφαρμογή της συνάρτησης πυρήνα δίνεται παρακάτω:

$$max \sum_{i=1}^{N} \pi_i K(X_i, X_i) - \sum_{i,j=1}^{N} \pi_i \pi_j K(X_i, X_j) \tag{5.26}$$

$$\text{δεδομένου ότι } \sum_{i=1}^{N} \pi_i = 1 \tag{5.27}$$

$$0 \leq \pi_i \leq \frac{1}{vn}, i = 1, ..., N, \tag{5.28}$$

5.4 Το διάγραμμα rk (*robust kernel − distance control chart*)

Παρακάτω παρουσιάζεται ένα νέο μη παραμετρικό πολυμεταβλητό διάγραμμα ελέγχου που βασίζεται στην απόσταση πυρήνα που ξεπερνά τους περιορισμούς της χρήσης της έννοιας της ταξινόμησης μίας τάξης που βασίζεται στις αρχές των διανυσμάτων υποστήριξης. Το διάγραμμα είναι μη παραμετρικό και αυτό δεν απαιτεί υποθέσεις όσο αφορά την πυκνότητα πιθανότητας και την απαίτηση μόνο "κανονικών" ή εντός ελέγχου δεδομένων για αποτελεσματική αναπαράσταση μίας εντός ελέγχου διεργασίας. Δίνει επίσης μία σαφή πρόβλεψη για την ενσωμάτωση οποιουδήποτε διαθέσιμου δεδομένου που βρίσκεται σε εκτός ελέγχου διαδικασία. Πειραματικές εκτιμήσεις σε μία ποικιλία δεδομένων βαθμολόγησης επιδόσεων προτείνουν το συγκεκριμένο διάγραμμα σαν αποτελεσματική διαδικασίας παρακολούθησης.

Η σταθερή υπόθεση πίσω από την πλειοψηφία των SPC και $MSPC$ μεθόδων είναι ότι οι μεταβλητές της διεργασίας ακολουθούν *Gaussian* κατανομή (Rose (1991)), μία αμφισβητήσιμη κατανομή σε πολλές βιομηχανικές διεργασίες (Polansky (2001)) και γενικά σε αυτοματοποιημένες διεργασίες (Chinnam and Kolarik (1992)). Οι Schilling and Nelson (1976) και άλλοι ερευνητές εξετάζουν τα αποτελέσματα της μη-κανονικότητας των ορίων ελέγχου και την απόδοση της χαρτογράφησης (*charting*). Για να εξαλειφθούν αυτά τα αποτελέσματα, προτείνονται μερικά διαγράμματα ελέγχου χωρίς κατανομή ή μη παραμετρικά, που βασίζονται στις διαδοχικές τάξεις των μετρήσεων του βάθους των δεδομένων (Liu and Singh (1993), Aradhye et al. (2001), Stoumbos and Reynolds (2001), Chakraborti et al. (2003), Messaoud et al. (2004)), αλλά η εξέλιξη και η παρουσίαση αυτών είναι αργή στο βιομηχανικό έλεγχο διεργασίας (Chakraborti et al. (2001)).

Διάφορα μη παραμετρικά διαγράμματα ελέγχου που βασίζονται στη θεωρία μάθησης και στις αρχές της αναγνώρισης μονοπατιών προτείνονται στη βιβλιογραφία. Για παράδειγμα, οι Cook and Chiu (1998) προτείνουν δίκτυα συναρτήσεων ακτινικής βάσης ($radial\ basis\ function - RBF\ networks$) για τον εντοπισμό αλλαγών σε συσχετισμένες διεργασίες κατασκευής, ο Chinnam (2002) πρότεινε τις μηχανές διανυσμάτων υποστήριξης για τον εντοπισμό αλλαγών σε συσχετισμένες και άλλες διεργασίες κατασκευής και οι Smith (1994) και Pugh (1991) μελέτησαν τα δίκτυα πολυστρωματικών αισθητήρων ($multi-layer\ perceptron - MLP$) για την εκτέλεση των διαγραμμάτων ελέγχου τύπου *Shewhart*, και όλα τα παραπάνω χαλαρώνουν την υπόθεση ότι οι μεταβλητές ακολουθούν *Gaussian* κατανομή. Όσο αυτές οι μέθοδοι δείχνουν επιτυχία στη χαλάρωση της υπόθεσης *Gaussian* κατανομής, τα θεμελιώδη όρια τα οποία προτείνουν το παρακάτω και πολλές άλλες μέθοδοι μηχανικής μάθησης στη βιβλιογραφία του ελέγχου διεργασίας, απορρίπτουν προβλήματα όπως αυτό της ταξινόμησης ή της αναγνώρισης μοτίβων, και έτσι είναι απολύτως αναγκαία τα δεδομένα παραδειγμάτων από όλες τις εκτός ελέγχου καταστάσεις που μας ενδιαφέρουν. Αυτός ο κρίσιμος περιορισμός για την εξασφάλιση παραδειγμάτων καταστάσεων από όλες τις καταστάσεις μπορεί να είναι δύσκολος, ακριβός, και ακόμα και αδύνατος. Ο δεύτερος περιορισμός είναι οτι δεν χρειάζεται καμία σαφής διάταξη για τις αλλαγές μεταξύ σφαλμάτων τύπου I και τύπου II. Οι περισσότερες από αυτές τις μεθόδους μηχανικής μάθησης καθιστούν αναγκαία τη μοντελοποίηση και την εκπαίδευση για κάθε συγκεκριμένο τύπο αποτυχίας. Ένα μοντέλο που αναπτύχθηκε για ένα συγκεκριμένο τύπο αποκλίνοντος συμβάντος (εκτός ελέγχου κατάσταση) δεν μπορεί απαραίτητα να δώσει ταξινόμηση ακριβείας για κάποιον άλλο τύπο αποκλίνοντος συμβάντος.

Το μη παραμετρικό διάγραμμα ελέγχου απόστασης πυρήνα που παρουσιάζεται παρακάτω, χρησιμοποιεί την ιδέα της ταξινόμησης μίας τάξης ή της καινοτομίας ανίχνευσης για να ξεπεράσει αυτούς τους περιορισμούς και υιοθετεί τις αρχές των μηχανών διανυσματικής υποστήριξης για να το κάνει αυτό. Το προτεινόμενο διάγραμμα ελέγχου βασίζεται στις αρχές του SVM για τους παρακάτω λόγους:

1. Είναι ένα σύστημα για αποτελεσματική εκπαίδευση μηχανών γραμμικής εκμάθησης στο χαρακτηριστικό χώρο που επηρεάζεται από τον πυρήνα,

2. ελέγχουν επιτυχώς την ευελιξία του χαρακτηριστικού χώρου που είναι επηρεασμένος από τον πυρήνα μέσω της γενικευμένης θεωρίας και

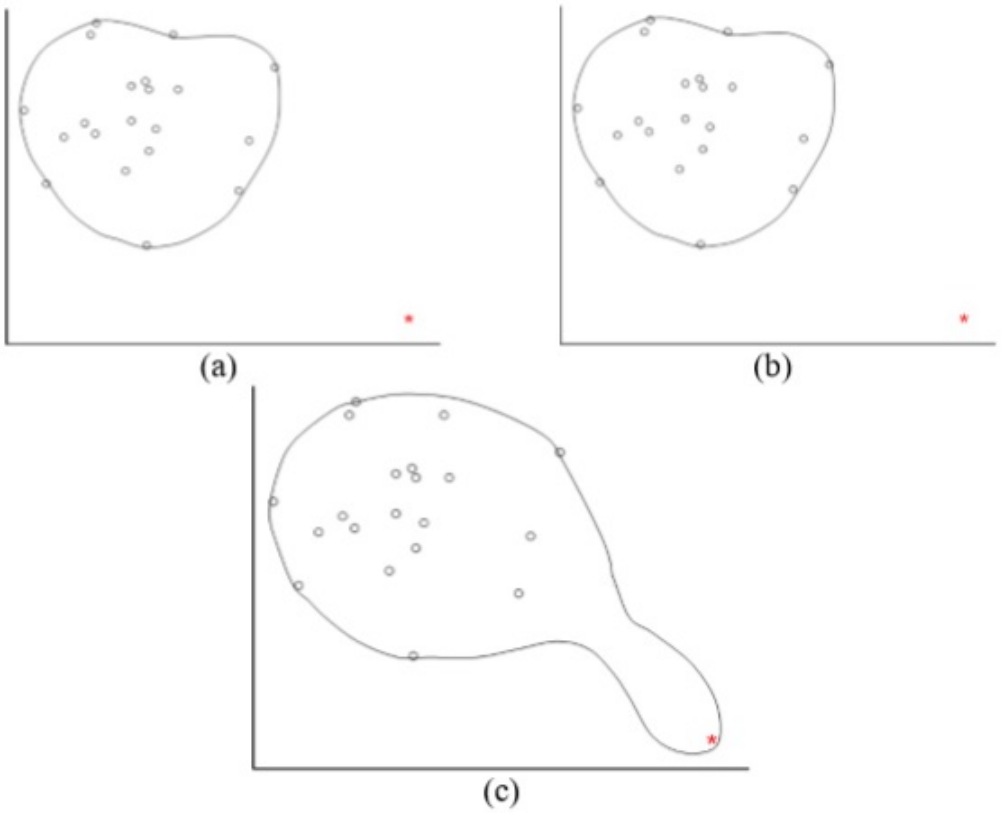
3. εκμεταλλεύονται την υπάρχουσα θεωρία βελτιστοποίησης για να το κάνουν αυτό.

Ένα σημαντικό χαρακτηριστικό των SVM συστημάτων είναι ότι όταν θέτουν σε λειτουργία την εκπαίδευση που προτείνεται από τη γενικευμένη θεωρία, πα-

ράγουν σποραδικές δυαδικές αναπαραστάσεις της υπόθεσης, με αποτέλεσμα πολύ αποδοτικούς αλγορίθμους. (Cristianini και Taylor (2000)). Αυτό συμβαίνει εξαιτίας των συνθηκών των Karush-Kuhn-Tucker (Kuhn και Tucker (1951)), (Mangasarian (1994)), στις οποίες βασίζεται η λύση και παίζουν πολύ σημαντικό ρόλο στην εκτέλεση και την ανάλυση τέτοιων μηχανών. Ένα άλλο σημαντικό χαρακτηριστικό της προσέγγισης με διανύσματα υποστήριξης είναι ότι με τις συνθήκες του Mercer στους πυρήνες (Mercer (1909)), τα αντίστοιχα προβλήματα βελτιστοποίησης είναι πολύπλοκα και έτσι δεν έχουν τοπικό ελάχιστο. Αυτό το γεγονός και ο μειωμένος αριθμός των μη μηδενικών παραμέτρων, σημειώνουν μία διάκριση μεταξύ αυτών των συστημάτων και άλλων αλγορίθμων μηχανικής μάθησης, όπως τα νευρονικά δίκτυα. Το τελικό αποτέλεσμα είναι ότι το προτεινόμενο διάγραμμα ελέγχου απόστασης πυρήνα είναι μη παραμετρικό και απαιτεί μόνο δεδομένα από την εντός ελέγχου κατάσταση, προβλέπει τη χρήση οποιουδήποτε διαθέσιμου δεδομένου από τις καταστάσεις εκτός ελέγχου και επιτρέπει αλλαγές μεταξύ των σφαλμάτων τύπου I και τύπου II. Υποστηρίζει μονομεταβλητές και πολυμεταβλητές διεργασίες και μπορεί να παρακολουθέι μαζί την τοποθεσία διεργασίας και τις πτυχές της διασποράς μέσω ενός μόνο διαγράμματος ελέγχου.

Το K διάγραμμα είναι ένας άλλος τύπος διαγραμμάτων ελέγχου απόστασης πυρήνα. Παρόλα αυτά ένας σημαντικός περιορισμός που προκύπτει με τη χρήση της $SVDD$, είναι ότι το K διάγραμμα των Sun and Tsung (2003) στερείται οποιασδήποτε ικανότητας να κάνει καλή διάκριση μεταξύ των ακραίων τιμών και των κανονικών δεδομένων στην ομάδα εκπαίδευσης. Πρόσθετα, δεν προβλέπει τη χρήση οποιουδήποτε διαθέσιμου δεδομένου από τις καταστάσεις εκτός ελέγχου, και τέλος δεν προσφέρει μια δομημένη μέθοδο για τη δημιουργία αλλαγών μεταξύ των σφαλμάτων τύπου I και II. Αυτό επίσης μπορεί να προκαλέσει μία φτωχή αναπαράσταση της εντός ελέγχου κατάστασης, αν τα διαθέσιμα δεδομένα για την αρχικοποίηση του διαγράμματος ελέγχου δεν είναι προεπεξεργασμένα για την εξάλειψη των ακραίων τιμών. Το προτεινόμενο διάγραμμα ελέγχου ξεπερνά αυτούς τους περιορισμούς με την ενσωμάτωση της $SVDD$ μεθόδου. Στο Σχήμα 5.3 παρουσιάζονται πολλά από αυτά τα πλεονεκτήματα του προτεινόμενου μοντέλου.

Αν κάποιος χρησιμοποιήσει δεδομένα από το Σχήμα 5.3(a) για να αρχικοποιήσει το K-διάγραμμα, δεν είναι εγγυημένο ότι το ακραίο σημείο θα αναγνωριστεί σαν ακραίο σημείο. Αντίθετα το διάγραμμα απόστασης πυρήνα που προτείνεται εδώ δίνει διάφορες επιλογές. Υποθέτοντας ότι το ακραίο σημείο δεν επισημαίνεται και παρουσιάζεται για την αρχικοποίηση του διαγράμματος, η προτεινόμενη μέθοδος αναγνωρίζει το σημείο σαν ακραίο, όπως στο Σχήμα 5.3(a). Αν το ακραίο σημείο επισημαίνεται σαν ακραίο πριν την αρχικοποίηση του διαγράμματος, η μέθοδος αναγνωρίζει την ετικέτα και το τοποθετεί εκτός



Σχήμα 5.4: Η ευελιξία του rk διαγράμματος.

του κανονικού ορίου, όπως στο Σχήμα 5.3(b). Αν για κάποιο λόγο κάποιος επιλέξει να χρησιμοποιήσει αυτό το σημείο σαν κανονικό η προτεινόμενη μέθοδος θα δεχτεί αυτόν τον περιορισμό και θα μεταχειριστεί το σημείο σαν να ήταν κανονικό και θα καθορίσει το κανονικό όριο της διεργασίας με το σημείο αυτό σαν σημείο του ορίου, όπως φαίνεται στο Σχήμα 5.3(c).

5.4.1 Το rk-διάγραμμα σαν ταξινομητής μίας κλάσης

Όπως αναφέρθηκε και νωρίτερα, το προτεινόμενο διάγραμμα ελέγχου χρησιμοποιεί την έννοια της ταξινόμησης μίας κλάσης για την παρακολούθηση διεργασιών, και για να το κάνει αυτό, μοντελοποιεί το όριο των δεδομένων της διεργασίας από μία κατάσταση εντός ελέγχου και αποφασίζει αν η διεργασία είναι εντός ή εκτός ελέγχου, αναλόγως με το αν η νέα παρατήρηση είναι εντός

ή εκτός του ορίου που υπάρχει στον χαρακτηριστικό χώρο.

Όπως πολλές $MSPC$ μέθοδοι, το rk διάγραμμα μοντελοποιεί από κοινού τις μετρήσεις της κεντρικής τάσης (*central tendecy*) και της διασποράς σε ένα πολυ-διάστατο χώρο. Όταν ο αριθμητικός μέσος και η τυπική απόκλιση θεωρούνται στατιστικές μετρήσεις για την παρακολούθηση της διεργασίας, η μέθοδος του rk διαγράμματος είναι μία γενική μέθοδος και μπορεί να ενσωματώσει οποιοδήποτε τύπο τοποθεσίας και διασποράς μετρήσεων. Η μόνη απαίτηση του rk διαγράμματος είναι ότι το διάνυσμα του δείγματος των στατιστικών μετρήσεων είναι πραγματικό, $\mathbf{x}_i \in \mathbf{R^d}$, όπου το d συμβολίζει τη διάσταση του διανύσματος. Ο αριθμός των διανυσμάτων υποστήριξης που είναι απαραίτητος για την αναπαράσταση της κανονικής ή της εντός ελέγχου κατάστασης της διεργασίας, θα αυξηθεί σαν συνάρτηση του d. Στην περίπτωση των πολυμεταβλητών διεργασιών, ο χαρακτηριστικός χώρος θα μπορούσε να είναι απλά ο χώρος μεταβλητών της διεργασίας ή μία μεταφορά αυτού, όπως για παράδειγμα ο χώρος των κύριων συνιστωσών. Στην περίπτωση των μονομεταβλητών διεργασιών, είναι απαραίτητο το μέγεθος των υποομάδων του δείγματος να είναι μεγαλύτερο από αυτό που διευκολύνει την εξαγωγή τουλάχιστον δύο χαρακτηριστικών, όπως ο μέσος και η τυπική απόκλιση. Ενώ τα rk διαγράμματα μπορούν θεωρητικά να αντιμετωπίσουν μονομεταβλητές και πολυμεταβλητές διεργασίες, και να παρακολουθούν ταυτόχρονα τις μετρήσεις τοποθεσίας και της διασποράς, δεν έχει απαραίτητα την ικανότητα να αναγνωρίζει τον τύπο του σφάλματος της διεργασίας, δηλαδή τις αλλαγές στην τοποθεσία και στη διασπορά.

Έτσι στην ανάπτυξη του προτεινόμενου διαγράμματος ελέγχου, υπάρχει μία απόκλιση από τον σχεδιασμό των συνηθισμένων SVM που σχεδιάζονται για δυαδική ταξινόμηση για την αναπαράσταση του ορίου δεδομένων μίας τάξης. Το rk διάγραμμα είναι εμπνευσμένο από την $SVDD$ (Tax and Duin (1999)) και τη $SVRM$ (*Support Vector Representation Machine*), (Yuan and Casasent (2003)) και δίνει τον ελάχιστο όγκο κλειστού σφαιρικού ορίου γύρω από τα δεδομένα της διεργασίας που είναι εντός ελέγχου, και αναπαριστάται από το κέντρο c και την ακτίνα r. Η ελαχιστοποίηση του όγκου γίνεται με την ελαχιστοποίηση του r^2, το οποίο αναπαριστά το κατασκευαστικό σφάλμα (Muller et al. (2001)):

$$minr^2 \tag{5.29}$$

$$\text{δεδομένου ότι: } \|\mathbf{x_i} - \mathbf{c}\|^2 \leq r^2, \forall i. \tag{5.30}$$

Ο περιορισμός αυτός δεν επιτρέπει σε οποιοδήποτε δεδομένο να βρεθεί εκτός της σφαίρας. Για να προβλεφθεί μία πιθανή ακραία τιμή εντός του δείγματος εκπαίδευσης, προτείνεται μία συνάρτηση ποινής (για δεδομένα εκτός της σφαίρας):

$$minr^2 + C\sum_i \xi_i \tag{5.31}$$

$$\text{δεδομένου ότι: } \|\mathbf{x_i} - \mathbf{c}\|^2 \leq r^2 + \xi_i,\ \xi_i \geq 0\ \forall i, \tag{5.32}$$

όπου το C είναι ο συντελεστής ποινής για κάθε ακραία τιμή (αναφέρεται και ως παράμετρος κανονικοποίησης) και ξ_i είναι η απόσταση μεταξύ της i-οστής παρατήρησης και της υπερσφαίρας. Το πρόβλημα βελτιστοποίησης δίνεται από τον τύπο 5.4 και μπορεί να λυθεί θεωρόντας τους πολλαπλασιαστές *Lagrange* σαν περιορισμούς:

$$L(R, O, \pi_i, \tau_i, \xi_i) = R^2 + C\sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \pi_i[R^2 + \xi_i - (X_i - O)^T(X_i - O)] - \sum_{i=1}^{N} \tau_i\xi_i,$$

όπου ξ_i και τ_i είναι οι πολλαπλασιαστές *Lagrange*, $\xi_i \geq 0$, $\tau_i \geq 0$. Για κάθε δεδομένο εκπαίδευσης X_i, καθορίζονται αντίστοιχα τ_i και ξ_i. Το L ελαχιστοποιείται θεωρώντας σταθερά τα R, O και π_i και μεγιστοποείται θεωρώντας σταθερά τα ξ_i και τ_i. Παραγωγίζοντας την 5.4 με σταθερά τα R, O και π_i και εξισώνοντας με το 0, παίρνουμε ότι:

$$O = \sum_i \tau_i \mathbf{x_i} \tag{5.33}$$

$$C - \tau_i - \xi_i = 0, \forall i \tag{5.34}$$

$$\sum_i \tau_i = 1, \tag{5.35}$$

και δεδομένου ότι $\xi_i \geq 0$, $\tau_i \geq 0$ ο περιορισμός 5.23 μπορεί να γραφεί ως

$$0 \leq \tau_i \leq C, \forall i. \tag{5.36}$$

Αντικαθιστώντας τις εξισώσεις 5.22, 5.23, 5.24 και 5.25 στην 5.4 έχουμε τις εξισώσεις τετραγωνικού προγραμματισμού:

$$max \sum_i \tau_i(X_iX_i) - \sum_{i,j} \tau_i\tau_j(X_iX_j) \tag{5.37}$$

$$\text{δεδομένου ότι } 0 \leq \tau_i \leq C\ \forall i, \sum_i \tau_i = 1. \tag{5.38}$$

Οπότε η διατύπωση του rk διαγράμματος ικανοποιεί τους περιορισμούς (KKT[1]) για την επίτευξη μίας βέλτιστης λύσης. Παρατηρώντας ότι $C = \tau_i + \xi_i$, αν ένας

[1]Karush-Kuhn-Tucker

από τους πολλαπλασιαστές πάρει την τιμή 0, ο άλλος παίρνει την τιμή C. Όταν ένα σημείο $\mathbf{x_i}$ είναι εντός της σφαίρας, το αντίστοιχο τ_i θα είναι ίσο με 0. Αν είναι εκτός της σφαίρας, δηλαδή $\pi_i > 0$, το ξ_i θα είναι 0, οπότε το τ_i θα είναι ίσο με C. Όταν ένα σημείο είναι στο όριο τα τ_i και ξ_i θα βρίσκονται στο διάστημα [0,C]. Η λύση τετραγωνικού προγραμματισμού συνήθως παράγει λίγα σημεία με μη μηδενικό τ_i, ή διανύσματα υποστήριξης. Ιδιαίτερο ενδιαφέρον παρουσιάζει το γεγονός ότι αυτά τα διανύσματα υποστήριξης μπορούν να παρουσιάσουν αποτελεσματικά τα δεδομένα στον υπόλοιπο χώρο. Έστω $\mathbf{S}^{SV} = \{x_i : \tau_i \neq 0\}$ η ομάδα των διανυσμάτων υποστήριξης.

Γενικά, είναι απίθανο η υπερσφαίρα να δώσει μία καλή αναπαράσταση του ορίου σε μία εντός ελέγχου διεργασία στον αρχικό χώρο εισόδου. Έτσι η 5.26 αντικαθιστάται από τη συνάρτηση πυρήνα, οδηγώντας στο ακόλουθο πρόβλημα τετραγωνικού προγραμματισμού:

$$max \sum_i \tau_i K(X_i, X_i) - \sum_{i,j} \tau_i \tau_j K(X_i, X_j) \tag{5.39}$$

$$\text{δεδομένου ότι } 0 \leq \tau_i \leq C \ \forall i, \sum_i \tau_i = 1 \tag{5.40}$$

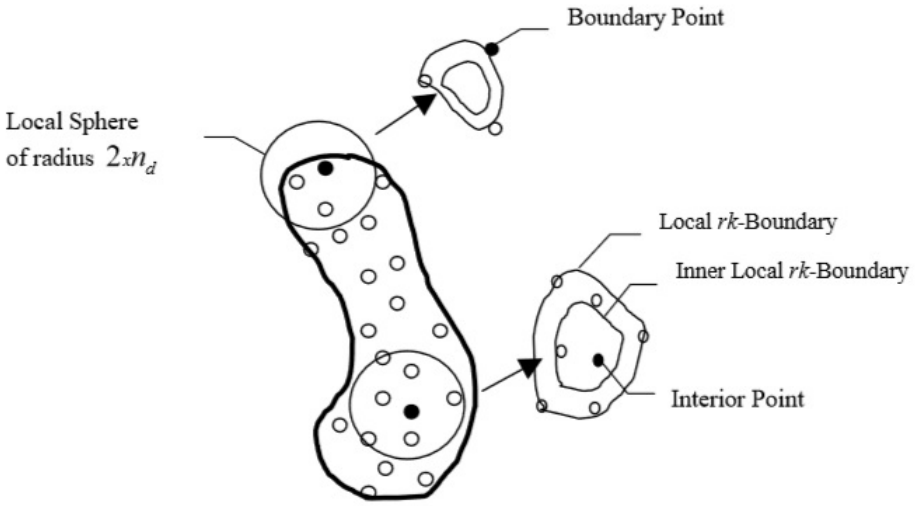
5.4.2 Ο Gaussian πυρήνας βελτιστοποίησης στο rk διάγραμμα

Χρησιμοποιείται ο *Gaussian* πυρήνας γιατί προσφέρει καλύτερη απόδοση σε σχέση με άλλους πυρήνες για τα προβλήματα ταξινόμησης μίας τάξης (Tax (2001)). Η συνάρτηση του *Gaussian* πυρήνα είναι:

$$K(\mathbf{x}, \mathbf{z}) = e^{-\|\mathbf{x}-\mathbf{z}\|^2/\sigma^2} \tag{5.41}$$

και το ζητούμενο είναι η βελτιστοποίηση της παραμέτρου σ. Αν το σ μπορεί να καθοριστεί από τον χρήστη, το rk διάγραμμα χρησιμοποιεί μία συγκεκριμένη διαδικασία για την επιλογή του σ (Tax and Duin (1999)).
Γενικά σε ένα σφαιρικό rk διάγραμμα, μικρότερες τιμές του σ παράγουν περισσότερα σημεία αναπαράστασης και μία μικρότερη υπερσφαίρα, ενώ μεγάλες τιμές δίνουν περισσότερα διανύσματα υποστήριξης και μεγαλύτερη υπερσφαίρα. Ο στόχος είναι να βρεθεί η τιμή του σ που δίνει αποτελέσματα της λίστας των διανυσμάτων υποστήριξης $\mathbf{S}^{SV}$ του ορίου του σφαιρικού rk διαγράμματος που να συμφωνούν με τη λίστα ορίου $\mathbf{S}^{BL}$, η οποία είναι αποτέλεσμα των τοπικών rk-ορίων. Στο Σχήμα 5.4 φαίνεται η λίστα ορίων.

Γενικά μικρότερες τιμές του σ έχουν ως αποτέλεσμα μία λίστα σφαιρικού διανύσματος υποστήριξης που είναι υπερσύνολο της λίστας ορίου με μερικά σημεία που δεν είναι μέρος της λίστας ορίου. Αντίθετα, μεγάλα σ έχουν σαν αποτέλεσμα μία λίστα σφαιρικού διανύσματος υποστήριξης που είναι υποσύνολο της λίστας ορίου. Συνεπώς, το rk διάγραμμα υπολογίζει την καταλληλότητα



Σχήμα 5.5: Ο καθορισμός της λίστας ορίου. Τα σημεία στο όριο θα απoορριφθούν από το εσωτερικό τοπικό rk-όριο.

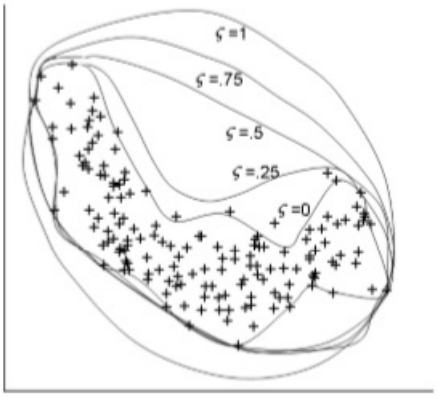
μίας τιμής σ χρησιμοποιώντας μία στρατηγική δύο σταδίων: της αποτελεσματικής αναπαράστασης και της συμπάγειας. Η αποτελεσματική αναπαράσταση πετυχαίνεται εξασφαλίζοντας ότι η λίστα σφαιρικού διανύσματος υποστήριξης ταιριάζει καλά στη λίστα ορίου. Η συμπάγεια αντίθετα δίνει έμφαση σε μία μικρότερη λίστα διανύσματος υποστήριξης, που βελτιώνει τη γενίκευση. Η συμπάγεια χειρίζεται μέσω μίας παραμέτρου που καθορίζεται από τον χρήστη, $0 \leq \varsigma \leq 1$. Όσο μεγαλύτερη είναι η τιμή του ς τόσο πιο συμπαγής είναι η λίστα διανύσματος υποστήριξης και τόσο μεγαλύτερο είναι το σφάλμα τύπου II, έχοντας σαν αποτέλεσμα μία μεγαλύτερη υπερσφαίρα.

Υπάρχει τυπικά μία τιμή του σ, που καθορίζεται από το σ_c και έχει σαν αποτέλεσμα μία σχεδόν τέλεια συμφωνία μεταξύ της λίστας διανύσματος υποστήριξης και της λίστας ορίου. Όσο το σ υπερβαίνει το σ_c, η λίστα διανύσματος υποστήριξης γίνεται μικρότερη. Η πραγματική τιμή του σ που χρησιμοποιείται στην καρασκευή του προτεινόμενου σφαιρικού rk διαγράμματος είναι:

$$\sigma = \sigma_c + \varsigma(\sigma_{max} - \sigma_c) \tag{5.42}$$

Το Σχήμα 5.6 αναπαριστά την επιρροή διαφορετικών επιπέδων συμπάγειας στην ποιότητα της αναπαράστασης χρησιμοποιώντας ένα παράδειγμα δεδομένων.

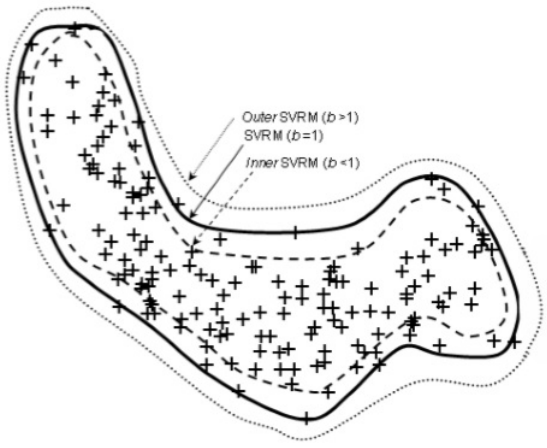
Το ενδότερο όριο παρέχει αποτελεσματική αναπαράσταση αλλά με 20 διανύσματα υποστήριξης, τα οποία όλα είναι στο $\mathbf{S}^{BL}$ ($\varsigma = 0$), ενώ το εξωτερικό πετυχαίνει συμπάγεια με μόλις 2 διανύσματα υποστήριξης ($\varsigma = 1$). Όσο οι παράμετροι σ_{min}, σ_{max} και σ_c υπολογίζονται εμπειρικά, η παράμετρος συμπάγειας



Σχήμα 5.6: Η επίδραση του ς στην αναπαράσταση ενός rk διαγράμματος.

ς και του ορίου του rk διαγράμματος πρέπει να προκαθοριστούν από τον χρήστη ή απαιτούνται αλλεπάλληλες δοκιμές ($\varsigma \approx 0.95$).

Αν υπολογιστεί η βέλτιστη τιμή του σ βάσει του βαθμού της συμπάγειας, μπορεί κανείς να κατασκευάσει τις αναπαραστάσεις του εσωτερικού και του εξωτερικού ορίου, αλλάζονται αντίστοιχα την ακτίνα της υπερσφαίρας του rk διαγράμματος. Το Σχήμα 5.7 παρουσιάζει αυτήν τη διαδικασία για τα ίδια δεδομένα με το Σχήμα 5.6. Είναι προφανές ότι όσο αλλάζει η ακτίνα, το γενικό γεωμετρικό σχήμα παραμένει ίδιο.



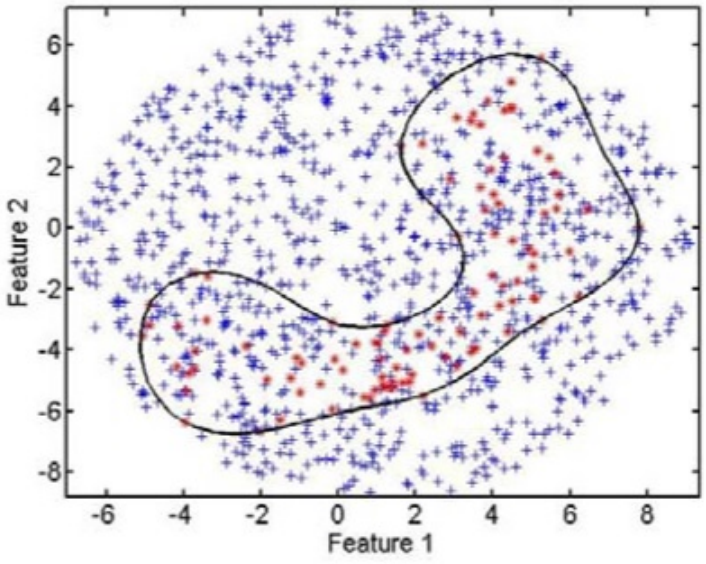
Σχήμα 5.7: Η επιρροή της ακτίνα της υπερσφαίρας του rk διαγράμματος στην αναπαράσταση του ορίου.

5.5 Το διάγραμμα ελέγχου που βασίζεται στον προσαρμοσμένο πυρήνα ($Adaptive\ Kernel - AK$)

Η κατασκευή ενός AK διαγράμματος απαιτεί δύο μέρη: την κατασκευή του $SVDD$ και τον καθορισμό των προσαρμοσμένων παραμέτρων.

5.5.1 Κατασκευή του *SVDD*

Αρχικά γίνεται η συλλογή της ομάδας δεδομένων μεγέθους N σε σταθερή κατάσταση. Μία υπόθεση που γίνεται στα προβλήματα ταξινόμησης μίας κλάσης είναι ότι τα δεδομένα είναι ανεξάρτητα και κατανεμημένα με τον ίδιο τρόπο. Οπότε μεταφέρουμε το σύνολο των δεδομένων με τη χρήση της PCA. Με αυτόν τον τρόπο μπορούμε να μειώσουμε τη συσχέτιση μεταξύ των μεταβλητών. Οι κύριες συνιστώσες (PC) που προκύπτουν από την PCA χρησιμοποιούνται στη συνέχεια για να καθοριστούν οι δύο παράμετροι της $SVDD$. Παρόλα αυτά, λόγω του ότι έχουμε δεδομένα μόνο μίας κλάσης, δεν μπορούμε να υπολογίσουμε το σφάλμα της ταξινόμησης που προκύπτει από τον καθορισμό των παραμέτρων της $SVDD$. Οι Tax and Duin (1999) πρότειναν μία μέθοδο για το πως να παράγει ομοιόμορφα τεχνιτές έκτροπες τιμές στην υπερσφαίρα (Σχήμα 5.8).



Σχήμα 5.8: Τεχνητά ομοιόμορφες ακραίες τιμές στην υπερσφαίρα.

Έτσι κάποιος μπορεί να έχει το σφάλμα τύπου I και II. Οι Ning and Tsung

(2013) χρησιμοποιούν την εξίσωση (5.17) για να επιλέξουν τις παραμέτρους της $SVDD$, αλλά είναι τόσο δύσκολο να κατανοηθούν στο βαθμό που θέλει ο χρήστης. Στη συνέχεια γίνεται αναδιατύπωση του μοντέλου:

$$min_{g,C}\beta \tag{5.43}$$

$$\text{δεδομένου ότι } \alpha \leq \alpha_0 \tag{5.44}$$

$$g > 0 \tag{5.45}$$

$$0 \leq C \leq 1 \tag{5.46}$$

όπου τα α και β είναι τα σφάλματα τύπου I και II, αντίστοιχα. Η χαρακτηριστική συνάρτηση (5.43) ελαχιστοποιεί το σφάλμα τύπου II. Το α_0 είναι το στοχευμένο σφάλμα τύπου I, το οποίο καθορίζεται από το χρήστη. Η εξίσωση (5.44) περιορίζει το σφάλμα τύπου I να είναι μικρότερο ή ίσο από το στόχο-σφάλμα τύπου I. Οι επόμενες δύο εξισώσεις προσδιορίζουν το εύρος των δύο παραμέτρων. Σε αυτό το μοντέλο, οι χρήστες πρέπει να καταλάβουν διαισθητικά τη σχέση μεταξύ των σφαλμάτων τύπου I και II.

Οι αλγόριθμοι αναζήτησης πλέγματος ($Grid\ Search$ - GS) είναι κατάλληλοι για τη λύση τέτοιων προβλημάτων, λόγω του ότι κάνουν πλήρη αναζήτηση σε ένα χώρο λύσης. Αφού το εύρος της παραμέτρου C είναι από 0 μέχρι 1, το εύρος του παραθύρου, g αλλάζει από 2^{-3} μέχρι 2^{10} σύμφωνα με την έρευνα των Ning and Tsung (2013). Πρώτα αναγνωρίζουμε τον αριθμό των τμημάτων, που χωρίζονται από το εύρος των δύο παραμέτρων και εκχωρούμε μία τιμή σε κάθε τμήμα. Στη συνέχεια συνδυάζουμε το κάθε τμήμα του εύρους του παραθύρου με το κάθε τμήμα συντελεστή ποινής σαν υπό-λύση, και χρησιμοποιούμε τις εξισώσεις (5.13-5.15) με κάθε υπό-λύση να κατασκευάζει μία υποψήφια τιμή για το βέλτιστο $SVDD$. Τελικά, κάποιος μπορεί να λύσει τις εξισώσεις (5.43-5.46) με όλες τις υποψήφιες τιμές για το $SVDD$ για να βρει το βέλτιστο $SVDD$. Το αποτέλεσμα χρησιμοποιείται για την κατασκευή του $SVDD$ ενός AK διαγράμματος.

5.5.2 Καθορισμός της παραμέτρου προσαρμογής (*adaptive parameter*)

Ένα AK διάγραμμα επιτρέπει στο διάστημα δειγματοληψίας του διαγράμματος να αλλάζει ανάλογα με τις προηγούμενες πληροφορίες. Σε αυτή τη μέθοδο, θεωρούμε δύο διαφορετικά διαστήματα δειγματοληψίας h_0 και h_1 και οι τιμές του δείγματος θα βρίσκονται σε τρεις περιοχές, οι οποίες χωρίζονται από τα όρια ελέγχου (UCL) και τα προειδοποιητικά όρια (WCL). Η περιοχή μεταξύ του μηδενός και του WCL είναι η ασφαλής περιοχή, δηλαδή η περιοχή στην οποία η διεργασία είναι ευσταθής. Η περιοχή μεταξύ του WCL και του UCL

είναι η προειδοποιητική περιοχή, όπου η διεργασία είναι κοντά στην εκτός ελέγχου κατάσταση. Η περιοχή εκτός του UCL είναι η περιοχή δράσης, όπου η διεργασία είναι εκτός ελέγχου. Η απόφαση αλλαγής μεταξύ των δύο διαφορετικών διαστημάτων δειγματοληψίας h_0 και h_1 εξαρτάται από την τοποθέτηση του πρότερου δείγματος. Αν το πρότερο δείγμα βρίσκεται στην ασφαλή περιοχή, θα χρησιμοποιηθεί το πιο μεγάλο διάστημα (h_0) στο καινούριο δείγμα. Αντίθετα, αν το πρότερο δείγμα είναι στην προειδοποιητική περιοχή, θα χρησιμοποιηθεί το πιο μικρό διάστημα (h_1). Αν το πρότερο δείγμα βρίσκεται στην περιοχή δράσης, η μέθοδος θεωρείται ότι βρίσκεται εκτός ελέγχου και ενεργοποιείται ένα σήμα. Η μέτρηση της απόδοσης του διαγράμματος AK εξαρτάται από τον καθορισμό των παραμέτρων προσαρμογής, συμπεριλαμβανομένων και των UCL, WCL, h_0 και h_1. Η μέτρηση της απόδοσης του διαγράμματος AK βασίζεται στο ATS, λόγω του ότι το διάστημα δειγματοληψίας ποικίλλει. Όταν η διεργασία είναι εντός ελέγχου, ο μέσος χρόνος σήματος (ATS_0) είναι μεγαλύτερος και η διεργασία καλύτερη και ο αριθμός εσφαλμένων συναγερμών μικρός. Επιπλέον, όταν η διεργασία είναι εκτός ελέγχου, το ATS_1 είναι μικρότερο, και διεργασία είναι καλύτερη. Έτσι ο χρόνος ανίχνευσης αλλαγών θα είναι μικρός. Και ο υπολογισμός των παραμέτρων δίνεται με τη λύση του παρακάτω προβλήματος βελτιστοποίησης:

$$min_{UCL,WCL,h_0,h_1} ATS_1 \tag{5.47}$$

$$\text{δεδομένου ότι } h_1 \geq h' \tag{5.48}$$

$$ATS_0 \geq ATS_0'. \tag{5.49}$$

Η αντικειμενική συνάρτηση (5.47) ελαχιστοποιεί το ATS_1. Η ανίσωση (5.48) περιορίζει την τιμή του h_1 να είναι μεγαλύτερη από το καθορισμένο διάστημα h' (το μικρότερο διάστημα δειγματοληψίας που μπορεί να έχει μία διεργασία). Και η ανίσωση (5.49) περιορίζει το ATS_0 να είναι μεγαλύτερο από την τιμή στόχο του σήματος, ATS_0', η οποία καθορίζεται από τους χρήστες. Παρακάτω δίνεται η εικόνα ενός AK διαγράμματος (Σχήμα 5.9).

Σχήμα 5.9: Ένα διάγραμμα AK.

5.5.3 Η απόδοση του διαγράμματος AK

Παρακάτω υπολογίζεται η απόδοση του προτεινόμενου διαγράμματος AK και συγκρίνεται με την απόδοση του T^2 διαγράμματος και του K διαγράμματος. Η σύγκριση μεταξύ αυτών των τριών διαγραμμάτων διεξάγεται για τις καταστάσεις πολυ-κανονικών κατανομών σε περιπτώσεις μικρών και μεγάλων διαστάσεων. Στην περίπτωση μικρής διάστασης $p = 3$ και στην περίπτωση μεγάλης διάστασης $p = 5$. Ο πίνακας διασποράς $\Sigma_0 = (\sigma_{ij})$ της πολυκανονικής κατανομής επιλέγεται να έχει $\sigma_{ii} = 1$ και $\sigma_{ij} = 0.5^{|i-j|}$ για $i, j = 1, 2, ..., p$. Σταθεροποιούμε την τιμή του α_0 στο 0.01 και θεωρούμε ως δ την μετατόπιση του μέσου της πρώτης παρατήρησης του σημείου δεδομένων. Επίσης παρουσιάζονται διαφορετικοί βαθμοί της μετατόπισης του μέσου της διεργασίας, $\delta = 0.25$, 0.5, 0.75, 1.00, 1.25, 1.50. Εδώ το εκτός ελέγχου ATS έχει επιλεχθεί σαν μέτρηση της απόδοσης του διαγράμματος ελέγχου, γιατί το διάστημα δειγματοληψίας είναι μεταβλητό. Και τα αποτελέσματα δίνονται μετά από 10000 επαναλήψεις.

Πίνακας 5.1: Πίνακας τιμών για $p = 3$

LOW DIMENSIONAL CASE

	ATS_1		
δ	T^2 chart	K-chart	AK-chart
0.25	36.14	22.74	14.52
0.50	7.52	1.28	1.13
1.00	1.41	1.00	1.00
1.25	1.03	1.00	1.00
1.50	1.00	1.00	1.00

Από τον Πίνακα 5.1 βγαίνει το συμπέρασμα ότι το K διάγραμμα και το AK διάγραμμα έχουν ικανοποιητική απόδοση και η διαφορά στην απόδοσή τους είναι μικρή στην περίπτωση μικρών διαστάσεων. Στην περίπτωση υψηλών διαστάσεων, όπως φαίνεται από τον Πίνακα 5.2, το διάγραμμα K έχει χειρότερη απόδοση από το διάγραμμα T^2 και δεν είναι ικανό να ανιχνεύει τις μετατοπίσεις ($\delta < 1.00$). Αλλά η διαφορά μεταξύ του AK διαγράμματος, του K διαγράμματος και του T^2 διαγράμματος είναι μικρή για μεγάλες μετατοπίσεις ($\delta > 1.25$). Σημειώνεται επίσης ότι το AK διάγραμμα έχει συνήθως καλύτερη απόδοση σε σχέση με τα άλλα δύο διαγράμματα και αυτό οφείλεται στο γεγονός ότι το AK διάγραμμα χρησιμοποιεί τις προηγούμενες πληροφορίες της διεργασίας για να ισχυροποιήσει την ανίχνευση μετατοπίσεων της διεργασίας.

Πίνακας 5.2: Πίνακας τιμών για $p = 5$

HIGH DIMENSIONAL CASE

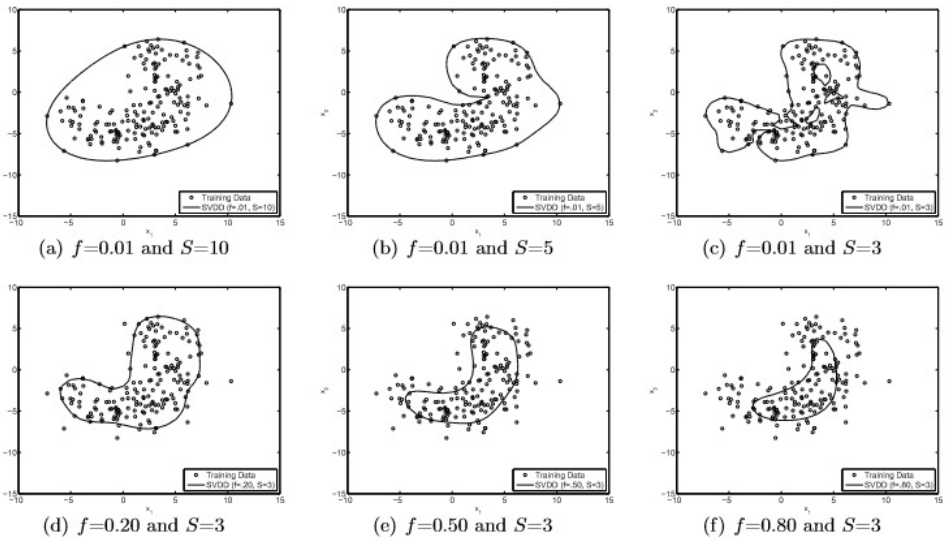
	ATS_1		
δ	T^2 chart	K-chart	AK-chart
0.25	74.7770	100.00	46.7679
0.50	16.8960	45.9560	11.6749
1.00	1.8650	6.1170	1.6279
1.25	1.00	3.21	1.01
1.50	1.00	1.00	1.00

Το AK διάγραμμα βελτιώνει την αδυναμία του K διαγράμματος, το οποίο δεν είναι ευαίσθητο στις μικρές μετατοπίσεις της διεργασίας. Επιπλέον, το διάγραμμα AK είναι πιο εύκολο να εφαρμοστεί από το K διάγραμμα, γιατί παρέχει ένα διαισθητικό τρόπο να εντοπίζει τη τιμή του σφάλματος τύπου I. Παρόλα αυτά, η κατασκευή του AK διαγράμματος απαιτεί τη γνώση της κατανομής των μετρήσεων γιατί ο καθορισμός των παραμέτρων προσαρμογής βασίζεται στη μέθοδο προσομοίωσης. Στην πραγματικότητα, αυτή η υπόθεση μερικές φορές δεν μπορεί ή είναι δύσκολο να ελεχθεί.

5.6 Το διάγραμμα ελέγχου μίας κλάσης (*one-class classification control chart*).

Τα διαγράμματα ελέγχου μίας τάξης SVM ($OC - SVM$) μπορούν να κατασκευαστούν από τον σχεδιασμό των στατιστικών D^2, τα οποία, όπως αναφέρθηκε, μετράνε την απόσταση μεταξύ των νέων παρατηρήσεων και του κέντρου της υπερσφαίρας. Τα όρια ελέγχου (R^2) των $OC - SVM$ διαγραμμάτων καθορίζονται από το f (ή το C). Με άλλα λόγια, ο βαθμός σφάλματος στα διαγράμματα $OC - SVM$ αυξάνεται με f. Μεγάλες τιμές της τιμής f έχουν σαν αποτέλεσμα μεγαλύτερη τιμή σφάλματος τύπου I, αφού ο αλγόριθμος χρησιμοποιεί λιγότερα δεδομένα μέσα στο όριο.

Το Σχήμα 5.11 αναπαριστά ένα T^2 διάγραμμα και δύο $OC - SVM$ διαγράμματα αντίστοιχα με τα όρια ελέγχου του Σχήματος 5.10. Σε αυτά τα σχήματα σχεδιάζονται τα στατιστικά που παρακολουθούνται από 400 δεδομένα που βρίσκονται στη Φάση II (τα πρώτα 360 βρίσκοντα εντός ελέγχου και τα υπόλοιπα 40 εκτός ελέγχου). Σημειώνεται επίσης ότι τα όρια ελέγχου για αυτά τα

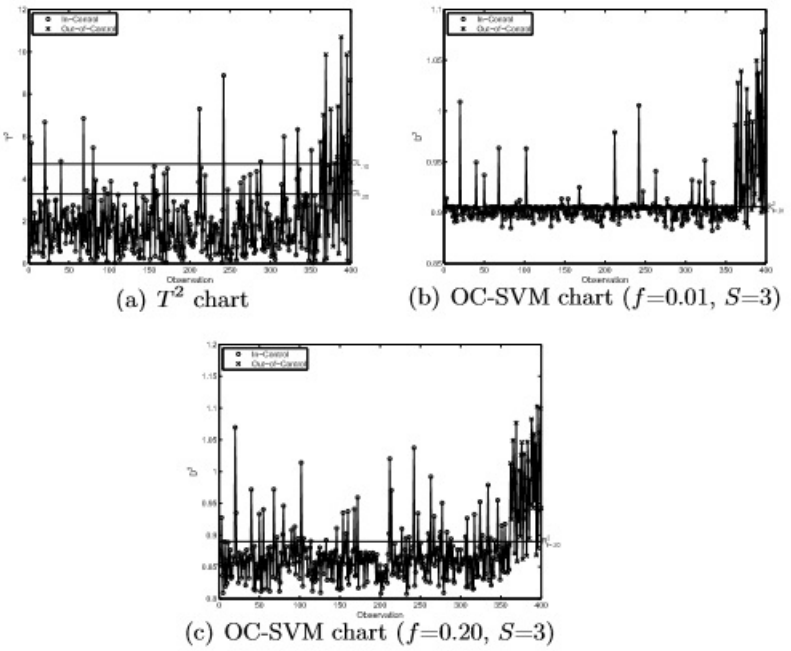


(a) f=0.01 and S=10 (b) f=0.01 and S=5 (c) f=0.01 and S=3

(d) f=0.20 and S=3 (e) f=0.50 and S=3 (f) f=0.80 and S=3

Σχήμα 5.10: Τα όρια ελέγχου του $SVDD$ που παρατηρούνται για διάφορες τιμές των παραμέτρων.

διαγράμματα ελέγχου έχουν καθοριστεί από 180 δεδομένα που βρίσκονται στη Φάση Ι. Στα $OC-SVM$ διαγράμματα, είναι ενδιαφέρον να παρατηρήσουμε ότι η τιμή του f που καθορίζεται από το χρήστη, δεν επηρεάζει μόνο τον καθορισμό των ορίων ελέγχου, αλλά και τον υπολογισμό του στατιστικού που παρακολουθείται. Παρατηρούνται δύο διαφορετικά διαγράμματα ελέγχου από την αλλαγή της τιμής f από 0.01 σε 0.20 (Σχήμα 5.10 c και d). Αυτό δείχνει ότι το f είναι ακατάλληλο για τον καθορισμό των ορίων ελέγχου στα $OC-SVM$ διαγράμματα. Ο περιορισμός αυτός μπορεί να εξηγηθεί από το Σχήμα 5.10, το οποίο δείχνει ότι παρατηρούνται εντελώς διαφορετικά όρια ελέγχου από την αλλαγή της τιμής του f από 0.01 σε 0.20. Σαν συνέπεια αυτού, μία παρατήρηση η οποία ανιχνεύθηκε σαν εκτός ελεγχου (εντός ελέγχου) μπορεί να μην ανιχνεύεται πλέον σαν εκτός ελέγχου (εντός ελέγχου) σαν αντίδραση στη χρήση διαφορετικών τιμών του f. Αντίθετα, τα T^2 διαγράμματα χρησιμοποιούν την ελεγχόμενη τιμή α, η οποία είναι ανεξάρτητη από το στατιστικό T^2. Έτσι, το ελλειψοειδές όριο του T^2 πάντα περιέχει περισσότερα εκτός ελέγχου σημεία και έχει σαν αποτέλεσμα μεγαλύτερη τιμή σφάλματος τύπου Ι με μεγαλύτερη τιμή του α (Σχήμα 5.10 b και c). Επιπλέον, οι ίδιες τιμές των στατιστικών που παρακολουθούνται σχεδιάζονται στο T^2 διάγραμμα, ανεξάρτητα από το α (Σχήμα 5.11 a).



(a) T^2 chart

(b) OC-SVM chart (f=0.01, S=3)

(c) OC-SVM chart (f=0.20, S=3)

Σχήμα 5.11: Τα διαγράμματα ελέγχου T^2 και $OC-SVM$ (a) T^2 διάγραμμα, (b) $OC-SVM$ διάγραμμα (f=0.01, s=3) και (c) $OC-SVM$ διάγραμμα (f =0.2, s =3).

5.6.1 Νέος σχεδιασμός του διαγράμματος $OC-SVM$.

Το διάγραμμα $OC-SVM$ που προτείνουν οι Sukchotrat et al. (2009) είναι το D^2 διάγραμμα. Τα όρια ελέγχου του D^2 διαγράμματος καθορίζονται και αυξάνονται βάσει μίας ποσοστιαίας τιμής, η οποία εκτιμάται από τη *bootstrap* μέθοδο.

Στα παραδοσιακά διαγράμματα ελέγχου, τα όρια ελέγχου καθορίζονται βάσει της κατανομής του στατιστικού που παρακολουθείται με την τιμή που καθορίζεται από το χρήστη. Αντίθετα, η κατανομή του στατιστικού που παρακολουθείται σε ένα D^2 διάγραμμα είναι άγνωστη λόγω της μη παραμετρικής της φύσης. Έτσι κατασκευάζουμε μία κατάλληλη μη παραμετρική διαδικασία για να σταθεροποιήσουμε το όριο ελέγχου. Πρώτα, βρίσκουμε τις παρατηρήσεις των D^2 τιμών της Φάσης ΙΙ μεγέθους N, μέσω του $SVDD$ αλγόριθμου. Στη συνέχεια με τα δείγματα B *bootstrap*, υπολογίζουμε τις ποσοστιαίες τιμές που μας ενδιαφέρουν από κάθε *bootstrap* δείγμα μεγέθους N αντικαθιστώντας τις παρατηρήσεις των τιμών D^2 Φάσης Ι. Τέλος, το όριο ελέγχου έχει καθοριστεί από το μέσο όρο των B ποσοστιαίων τιμών.

Παρακάτω περιγράφεται η διαδικασία *bootstrap* που καθορίζει τα όρια ελέγχου του D^2 διαγράμματος.

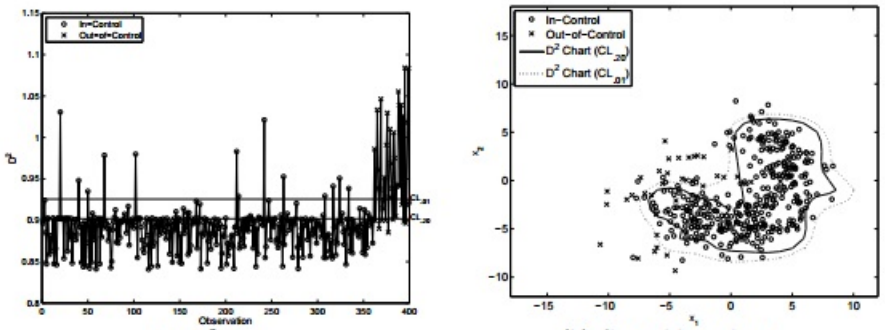
1. Υπολογίζουμε το στατιστικό D^2 για τις παρατηρήσεις της Φάσης Ι, μεγέθους N από τον τύπο (5.16) και παράγουμε B ανεξάρτητα (*bootstrap*) δείγματα. Έστω D^2_{j1}, D^2_{j2}, ... D^2_{jN} είναι η ακολουθία των N D^2 στατιστικών συναρτήσεων για το $j-$οστό δείγμα.

2. Για κάθε *bootstrap* δείγμα, δοθέντος ενός προκαθορισμένου από το χρήστη a $(0 < a \leq 1)$ και τις διατεταγμένες τιμές D^2 $(D^2_{j(1)} < D^2_{j(2)} < ... < D^2_{j(N)})$, όπου $D^2_{j(i)}$ είναι η i-οστή μεγαλύτερη από τις N D^2 τιμές του j-οστού *bootstrap* δείγματος, όπου i είναι ένας στρογγυλλοποιημένος αριθμός του $N * \alpha$.

3. Υπολογίζουμε το όριο ελέγχου (CL) παίρνοντας το μέσο από τις $i-$οστές μεγαλύτερες τιμές σε κάθε ένα από τα B *bootstrap* δείγματα: $CL = \sum_{j=1}^{B} D^2_{j(i)}/B$

4. Παρακολουθούμε τις παρατηρήσεις Φάσης ΙΙ: Δηλώνουμε τις παρατηρήσεις εκτός ελέγχου αν οι αντίστοιχες D^2 τιμές υπερβαίνουν το όριο ελέγχου.

Στο Σχήμα 5.12 παρουσιάζεται το διάγραμμα D^2 και τα αντίστοιχα όρια. Στο D^2 διάγραμμα ελέγχου, χρησιμοποιούνται 180 εντός ελέγχου παρατηρήσεις για να εκτιμηθούν τα όρια ελέγχου (βασισμένα στο *bootstrap* 99οστό και 80οστό ποσοστημόρια των D^2 στατιστικών) και σχεδιάζονται 400 παρατηρήσεις D^2 στατιστικών από τη Φάση ΙΙ. Το Σχήμα 5.12 (b) δείχνει τα αντίστοιχα όρια ελέγχου που δημιουργούνται από το D^2 διάγραμμα του Σχήματος 5.12 (a). Είναι προφανές ότι όσο αυξάνεται η τιμή του α από το 0.01 στο 0.20, ανιχνεύονται περισσότερες εκτός ελέγχου παρατηρήσεις.



(a) D^2 chart (b) Control boundaries

Σχήμα 5.12: Το D^2 διάγραμμα και το αντίστοιχο όριο ελέγχου.

5.7 Το διάγραμμα ελέγχου που βασίζεται στο *kNNDD*

Ο αλγόριθμος $SVDD$ περιέχει το πρόβλημα βελτιστοποίησης, το οποίο απαιτεί μεγάλο υπολογιστικό φορτίο κατά τη διαδικασία εκπαίδευσης. Ο $SVDD$ αλγόριθμος χρειάζεται περίπου 4.06 ώρες σε μία από τις μηχανές για να εκπαιδεύσει το μοντέλο χρησιμοποιώντας 4000 διδιάστατες παρατηρήσεις. Λογω αυτού, τα διαγράμματα δεν θα είναι αποτελεσματικά σε μία διεργασία που χρειάζεται συχνές επανεκπαιδεύσεις. Για να αντιμετωπιστεί αυτό το υπολογιστικό βάρος, προτάθηκε (Sukchotrat et al. (2009)) ένα νέο διάγραμμα ελέγχου που βασίζεται στην ταξινόμηση μίας κλάσης ($one-class\ classification$), το οποίο καλείται K^2 διάγραμμα. Ο αλγόριθμος που χρησιμοποιείται σε ένα K^2 διάγραμμα απαιτεί 5.42 δευτερόλεπτα (στην ίδια μηχανή με του $SVDD$ αλγορίθμου) για να εκπαιδεύσει 4000 διμεταβλητές παρατηρήσεις. Το διάγραμμα K^2 βασίζεται στη μέθοδο περιγραφής των δεδομένων των k κοντινότερων γειτόνων που λύνει προβλήματα ταξινόμησης μίας κλάσης, εκτιμώντας την τοπική πυκνότητα των δεδομένων χρησιμοποιώντας τον αλγόριθμο των κοντινότερων γειτόνων.

Ο αλγόριθμος kNNDD:

Έστω ότι $\mathrm{NN}_i(z)$ είναι ο i-οστός κοντινότερος γείτονας ενός σημείου z, ο οποίος πρέπει να ταξινομηθεί. Επίσης έστω V ο όγκος της υπερσφαίρας που περιέχει τις i κοντινότερες παρατηρήσεις-γείτονες. Έστω N το μέγεθος της ομάδας εκπαίδευσης. Η τοπική πυκνότητα του z καθορίζεται από τον τύπο:

$$d(z) = \frac{i/N}{V\|z - \mathrm{NN}_i(z)\|} \tag{5.50}$$

Παρόμοια η τοπική πυκνότητα του $\mathrm{NN}(z)$ καθορίζεται από τον τύπο:

$$d(\mathrm{NN}_i(z)) = \frac{i/N}{V\|\mathrm{NN}_i(z) - \mathrm{NN}_i(\mathrm{NN}_i(z))\|} \tag{5.51}$$

όπου $\mathrm{NN}_i(\mathrm{NN}_i(z))$ είναι ο i-οστός κοντινότερος γείτοντας του $\mathrm{NN}_i(z)$ στην ίδια ομάδα εκπαίδευσης. Ο αλγόριθμος $kNNDD$ ταξινομεί το z σαν την κλάση-στόχο όταν ο λόγος της τοπικής πυκνότητας του z προς την τοπική πυκνότητα του $\mathrm{NN}_i(z)$ είναι μεγαλύτερος ή ίσος με 1:

$$\frac{d(z)}{d(\mathrm{NN}_i(z))} = \frac{\|\mathrm{NN}_i(z) - \mathrm{NN}_i(\mathrm{NN}_i(z))\|}{\|\mathrm{z} - \mathrm{NN}_i(\mathrm{NN}_i(z))\|} \geq 1 \tag{5.52}$$

Για να γίνει πιο ακριβής ο αλγόριθμος, θεωρούμε το μέσο των k αποστάσεων:

$$\frac{\sum_{i=1}^{k} \|\mathrm{NN}_i(z) - \mathrm{NN}_i(\mathrm{NN}_i(z))\|}{\sum_{i=1}^{k} \|\mathrm{z} - \mathrm{NN}_i(\mathrm{NN}_i(z))\|} \geq 1. \tag{5.53}$$

Στο Σχήμα 5.13 φαίνονται τα όρια ελέγχου που ορίζονται από το $kNNDD$ με δύο διαφορετικές τιμές του k. Το όριο απόφασης για $k = 30$ είναι αρκετά ομαλό σε σύγκριση με το όριο ελέγχου για $k = 2$. Μία έρευνα υποστηρίζει ότι το κατάλληλο εύρος του k για τον $kNNDD$ αλγόριθμος είναι μεταξύ 10 και 50. (Breunig et al. (2000))

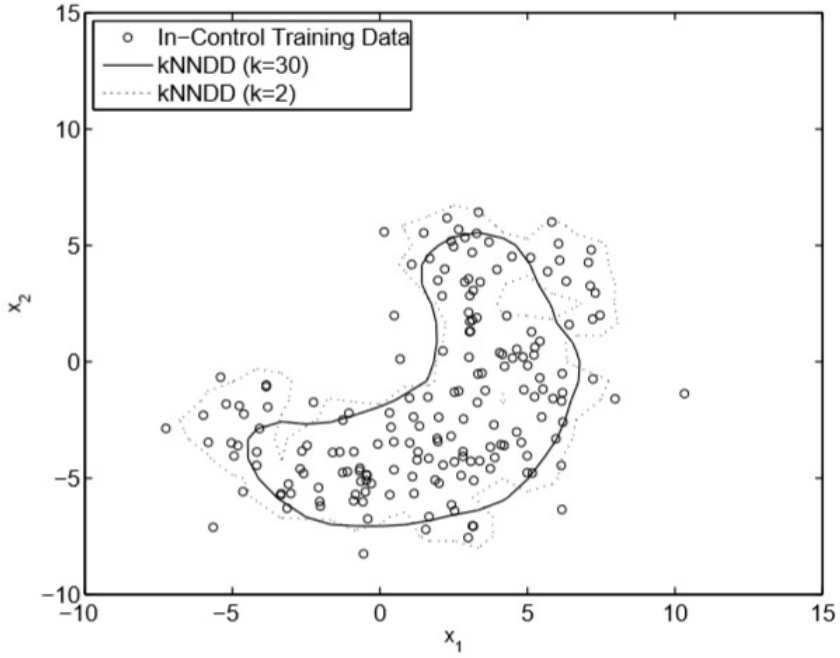
5.8 Τα διαγράμματα K^2

Για την κατασκευή των K^2 διαγραμμάτων (Sukchotrat et al. (2009)), υπολογίζεται η μέση απόσταση μεταξύ του z και των k κοντινότερων παρατηρήσεων:

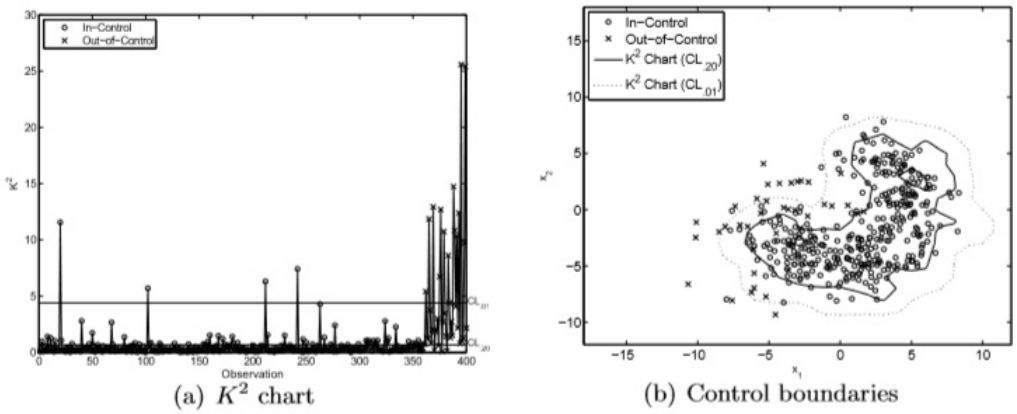
$$K^2 = \frac{\sum_{i=1}^{k} \|z - \mathrm{NN}_i(z)\|}{k} \tag{5.54}$$

Στη συνέχεια οι τιμές του K^2 χρησιμοποιούνται σαν στατιστικές συναρτήσεις παρακολούθησης. Τα όρια ελέγχου ενός K^2 διαγράμματος εξασφαλίζονται από την διαδικασία *bootstrap percentile* όπως και στα D^2 διαγράμματα.

Στο Σχήμα 5.14 παρουσιάζεται το διάγραμμα K^2 ($k = 30$) και τα αντίστοιχα όρια. Δύο διαφορετικά όρια ελέγχου υπόλογίστηκαν από τα εκτιμημένα ποσοστημόρια (99-οστό και 80-οστό) από 5000 *bootstrap* δείγματα των 180 K^2 στατιστικών. Για την παρακολούθηση των παρατηρήσεων της Φάσης II,



Σχήμα 5.13: Τα όρια ελέγχου του $kNNDD$ για διάφορα k.



(a) K^2 chart (b) Control boundaries

Σχήμα 5.14: Το K^2 διάγραμμα και τα αντίστοιχα όρια ελέγχου.

σχεδιάζεται η τιμή του K^2 για κάθε παρατήρηση της Φάσης ΙΙ. Από το δεύτερο διάγραμμα 5.14(b) φαίνεται ότι όσο η τιμή του a αυξάνεται, το διάγραμμα ελέγχου (δύο διαστάσεων) γίνεται πιο ευαίσθητο.

6 Εφαρμογές

6.1 Πολυμεταβλητά διαγράμματα ελέγχου χ^2, T^2, $MCUSUM$ και $MEWMA$

Παρακάτω γίνεται η μελέτη 3 ποιοτικών χαρακτηριστικών (εσωτερική διάμετρος του σωλήνα, πάχος και μήκος), ενός συγκεκριμένου σωλήνα ινών άνθρακα. Η ομάδα δεδομένων αποτελείται από 30 δείγματα μεγέθους 8 (τα οποία φαίνονται στον Πίνακα 6.1) και η διεργασία θεωρείται σταθερή.

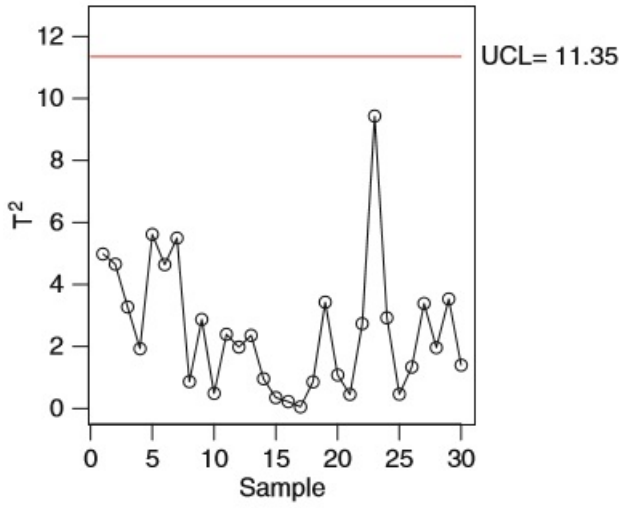
Πίνακας 6.1: Τα δεδομένα των ινών άνθρακα.

	Subgroup mean			Variance (×100)			Covariance (×100)			
Sample	Inner (X_1)	Thickness (X_2)	Length (X_3)	S_{21}	S_{22}	S_{23}	S_{12}	S_{13}	S_{23}	T^2
1	1.03	1.08	50.16	0.15	1.19	2.98	−0.08	0.40	−0.42	4.99
2	0.97	0.95	49.92	0.40	1.01	6.38	0.16	1.03	0.25	4.66
3	1.01	1.05	50.14	0.17	0.99	1.96	0.31	0.38	0.81	3.28
4	1.00	1.05	49.91	0.26	1.14	5.73	0.36	−0.29	0.82	1.93
5	0.96	1.00	49.83	0.43	3.25	12.37	0.92	1.91	4.61	5.62
6	1.03	1.07	50.05	0.30	1.52	1.69	0.43	0.30	−0.48	4.64
7	0.96	1.02	49.95	0.17	0.58	4.34	0.03	0.34	0.74	5.50
8	1.00	1.02	50.02	0.17	0.79	6.16	0.11	−0.11	0.32	0.87
9	1.00	1.10	50.03	0.20	1.43	1.87	0.34	−0.07	−0.88	2.87
10	0.99	1.02	50.00	0.13	0.53	6.58	0.11	0.18	−0.13	0.49
11	1.01	1.10	50.01	0.18	1.31	3.41	0.11	0.19	0.36	2.40
12	1.02	1.07	49.99	0.24	0.81	3.41	0.05	0.70	0.67	1.98
13	0.97	1.00	49.96	0.48	2.36	17.72	0.98	2.44	5.17	2.36
14	1.01	1.05	50.04	0.13	1.08	7.20	0.18	0.16	1.98	0.96
15	1.00	1.06	50.02	0.24	1.14	7.80	0.26	1.01	0.43	0.35
16	1.00	1.03	49.99	0.39	1.66	3.69	0.71	0.98	1.27	0.22
17	1.00	1.04	49.99	0.10	1.27	7.71	0.00	0.44	−2.09	0.05
18	0.98	1.00	49.94	0.18	1.56	5.40	0.26	0.85	2.22	0.86
19	0.98	0.96	49.93	0.24	1.61	5.68	0.55	0.18	0.55	3.43
20	1.01	1.07	50.02	0.37	2.55	4.91	0.64	1.16	3.33	1.08
21	0.98	1.03	49.96	0.28	0.39	7.21	0.15	1.39	0.64	0.45
22	0.99	1.04	50.07	0.23	2.46	8.24	0.60	0.70	1.74	2.74
23	0.95	0.92	49.86	0.41	1.82	2.69	0.73	0.40	0.32	9.43
24	1.00	1.09	50.05	0.15	0.75	9.27	0.12	0.69	−0.29	2.93
25	0.99	1.01	49.96	0.51	1.87	7.08	0.56	1.56	1.63	0.46
26	0.99	1.02	49.89	0.12	0.75	7.04	0.19	0.59	1.34	1.34
27	0.99	1.03	49.84	0.24	3.80	7.47	0.72	0.87	2.20	3.39
28	1.01	1.04	49.97	0.06	0.80	2.46	0.14	0.08	0.05	1.97
29	1.03	1.10	50.07	0.19	1.29	2.38	0.43	0.36	0.72	3.54
30	1.01	1.08	49.97	0.33	1.75	6.78	0.69	1.27	2.73	1.40

Παρακάτω δίνονται το διάνυσμα του μέσου, ο πίνακας συνδιασποράς και ο πίνακας συσχέτισης

$$\bar{x} = \begin{bmatrix} 0.99 \\ 1.04 \\ 49.98 \end{bmatrix}, S \times 100 = \begin{bmatrix} 0.25 & 0.36 & 0.67 \\ 0.36 & 1.45 & 1.02 \\ 0.67 & 1.02 & 5.92 \end{bmatrix}, r = \begin{bmatrix} 1 & 0.63 & 0.57 \\ 0.63 & 1 & 0.38 \\ 0.57 & 0.38 & 1 \end{bmatrix}$$

Είναι προφανές ότι υπάρχει συσχέτιση μεταξύ των μεταβλητών, που γίνεται σημαντική μεταξύ της εσωτερικής διαμέτρου και των άλλων.
Το στατιστικό T^2 παίρνει την τιμή $T^2 = 4.99$ για το πρώτο δείγμα και το άνω όριο ελέγχου $UCL = 11.35$. Στο Σχήμα 6.1 φαίνεται το T^2 διάγραμμα ελέγχου για το συγκεκριμένο δείγμα δεδομένων.

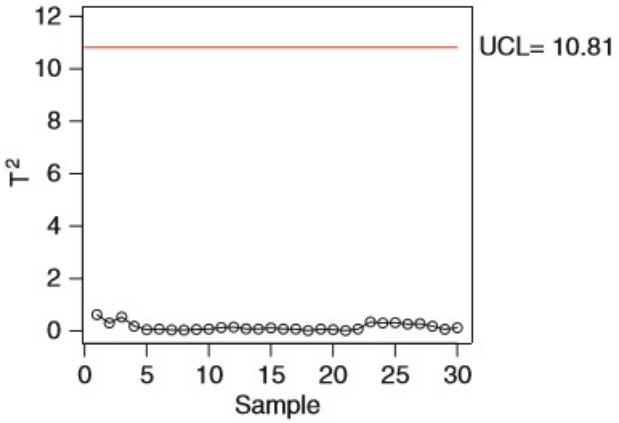


Σχήμα 6.1: Το T^2 διάγραμμα ελέγχου.

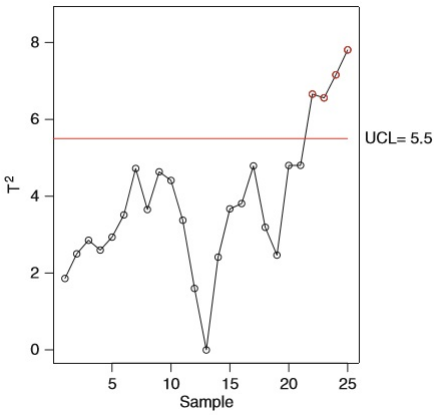
Παρατηρούμε ότι, δεν υπάρχουν σημεία που να ξεπερνούν το άνω όριο, οπότε η διεργασία είναι εντός στατιστικού ελέγχου.
Το διάγραμμα $MEWMA$ έχει το στατιστικό T^2=$Z_i' \sum_{Z_i}^{-1} Z_i$, όπου Z_i=$\lambda X_i + (1-\lambda)X_{i-1}$ με $Z_0 = 0$ και $\lambda = 0.1$. Το διάγραμμα $MEWMA$ για $\lambda = 0.1$, $ARL = 200$, $p = 3$ και UCL=10.81 δίνεται στο Σχήμα 6.2.

Στη συνέχεια σχεδιάζεται το $MCUSUM$ διάγραμμα με τη μέθοδο που προτείνουν οι Pignatiello και Runger (1990), που δίνει την καλύτερη απόδοση. T_i^2=$max\{0, [S_i'(\frac{\Sigma}{n})^{-1}S_i]^{1/2-kn_i}\}$, όπου $S_i = \sum_{j=i-n_i+1}^{i}(\bar{X}_i - \mu_0)$ και $n_i = n_{i-1}+1$, αν $T_{i-1}^2 > 0$ αλλιώς $n_i = 1$. Τέλος, έχουμε $UCL = h$. Στη συγκεκριμένη περίπτωση $n_i = 1$, οπότε S_1=$\{\begin{bmatrix} 1.01 & 1.07 & 49.88 \end{bmatrix} - \begin{bmatrix} 0.99 & 1.04 & 49.98 \end{bmatrix}\}$ και $T_i^2 = 1.86$.

Στο Σχήμα 6.3 παρατηρούνται σήματα εκτός ελέγχου.



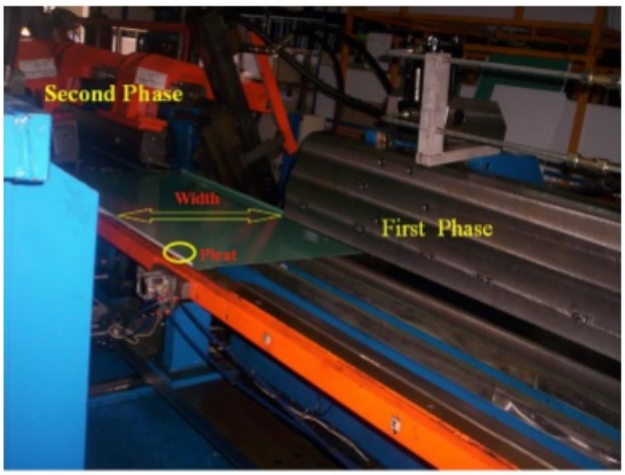
Σχήμα 6.2: Το $MEWMA$ διάγραμμα ελέγχου.



Σχήμα 6.3: Το $MCUSUM$ διάγραμμα ελέγχου που προτείνεται από τους Pignatiello και Runger.

6.2 Μία πραγματική μελέτη

Η συγκεκριμένη μελέτη (Gani et al. (2008)) ασχολείται με μία βιομηχανική εταιρεία που ειδικεύεται στην κατασκευή οικιακών ηλεκτρικών συσκευών, που έχει σαν κύριο προϊόν τα ψυγεία. Συγκεκριμένα ελέγχεται ένα φύλλο μετάλλου, όπου παρακολουθούνται ταυτόχρονα τέσσερα ποιοτικά χαρακτηριστικά. Αυτά είναι το πλάτος τύπου 1, το πλάτος τύπου 2, το πάχος τύπου 1 και το πάχος τύπου 2. Στο Σχήμα 6.4 αναπαριστάται η κατασκευή του φύλλου μετάλλου της διεργασίας. Η προετοιμασία του φύλλου μετάλλου για το ψυγείο γίνεται σε δύο βήματα, όπως φαίνεται στο Σχήμα 6.4. Στο πρώτο βήμα, το φύλλο πλισσάρεται (*pleated*) και τα δύο πάχη που καλούνται *pleat* I και *width* I, λαμ-



Σχήμα 6.4: Η κατασκευή ενός φύλλου μετάλλου.

βάνονται, αντίστοιχα. Στο δεύτερο βήμα, το φύλλο πλισσάρεται ξανά και δίνει δύο άλλα πάχη που καλούνται *pleat* II και *width* II, αντίστοιχα. Με αυτόν τον τρόπο, μετράμε τέσσερα ποιοτικά χαρακτηριστικά, εκφρασμένα σε *millimetre*. Παίρνουμε ένα δείγμα από 134 παρατηρήσεις, με συχνότητα δειγματοληψίας 30 λεπτά.

Η κατασκευή του K διαγράμματος απαιτεί τρία σημαντικά βήματα.

- Βήμα 1: Το σύνολο δεδομένων αναλύεται με τη χρήση της PCA για να μην υπάρχει πρόβλημα συσχέτισης μεταξύ των μεταβλητών.
- Βήμα 2: Οι κύριες συνιστώσες που προκύπτουν από το Βήμα 1, ταξινομούνται σε μία κλάση με τη χρήση της $SVDD$ μεθόδου. Σε αυτό το βήμα, η μία κλάση θα βελτιωθεί με την επιλογή βέλτιστων τιμών για τις παραμέτρους της $SVDD$, όπως την απόρριψη ορίων και του πλάτους. Το πρώτο κριτήριο, στο οποίο βασίζεται η βελτιστοποίηση, είναι ο αριθμός των διανυσμάτων υποστήριξης, τα οποία είναι ένας σημαντικός δείκτης για την κατασκευή μίας καλής κλάσης-στόχου. Συγκεκριμένα, η συνεισφορά των διανυσμάτων υποστήριξης στον υπολογισμό των σφαλμάτων στην κλάση-στόχο καθορίζεται από:

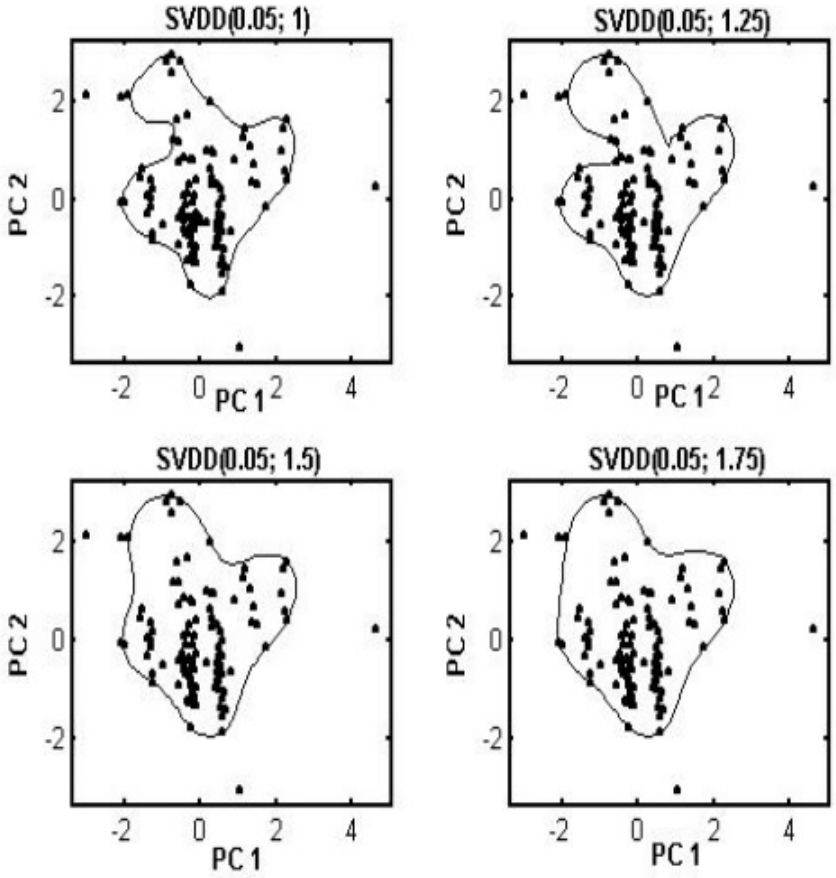
$$E[P(error)] = \frac{Nb.SV}{N}, \tag{6.1}$$

όπου $Nb.SV$ είναι ο αριθμός των διανυσμάτων υποστήριξης και N είναι ο αριθμός των παρατηρήσεων.

Το δεύτερο κριτήριο είναι ο όγκος της υπερσφαίρας. Είναι ένας σημαντικός δείκτης γιατί το αντικείμενο της $SVDD$ μεθόδου είναι η ανακάλυψη των περισσότερων δεδομένων με τον ελάχιστο όγκο υπερσφαίρας. Οι πρώτες 130 παρατηρήσεις χρησιμοποιούνται για την κατασκευή της κλάσης-στόχου. Στην πραγματικότητα, για να βρούμε μία καλή περιγραφή των δεδομένων με τη χρήση της $SVDD$, το μέγεθος των δεδομένων πρέπει να είναι αρκετά μεγάλο για να πάρουμε μία ακριβή μέτρηση.

- Βήμα 3: Η βέλτιστη κλάση που έχει βρεθεί στο βήμα 2, χρησιμοποιείται για το σχεδιασμό του K διαγράμματος, υπολογίζοντας το όριο ελέγχου βάσει των διανυσμάτων υποστήριξης που καθορίστηκαν από την $SVDD$ μέθοδο. Σε αυτό το βήμα, οι τελευταίες 4 παρατηρήσεις χρησιμοποιούνται για την ανίχνευση εκτός ελέγχου σημάτων. Η κατασκευή την κλάσης απαιτεί δύο παραμέτρους, τη συνάρτηση απόρριψης f και το εύρος (σ) του RBF ($Radial\ Base\ Function$). Στην πραγματικότητα, η f θεωρείται και η συνάρτηση των δεδομένων που θα απορριφθούν από την κλάση-στόχο. Διαφορετικά η f αναπαριστά το σφάλμα τύπου Ι στην ταξινόμηση μίας κλάσης. Το εύρος της RBF καθορίζει το σχήμα της κλάσης που θα κατασκευαστεί. Στην συγκεκριμένη μελέτη (Gani et al. (2008)), η τιμή f=0.05 χρησιμοποιείται για την κατασκευή της κλάσης-στόχου. Η ίδια τιμή χρησιμοποιείται και για τη συγκριτική μελέτη, όπου κατασκευάζεται το T^2 διάγραμμα με το σφάλμα τύπου Ι να ισούται με 0.05. Για f=0.05, χρησιμοποιούνται 4 τιμές για το σ, δηλαδή σ=1, 1.25, 1.5, 1.75. Οι 4 τελευταίες παρατηρήσεις χρησιμοποιούνται για την ανίχνευση ακραίων τιμών. Όπως φαίνεται στο Σχήμα 6.5, παρατηρούνται διαφορετικοί τύποι κλάσεων. Για κάθε κλάση ανιχνεύονται 4 ακραίες τιμές. Για σ=1, παρατηρούμε μια κλάση με 22 διανύσματα υποστήριξης. Παρόλα αυτά για σ=1.25 χρησιμοποιούνται μόνο 18 διανύσματα υποστήριξης για να κατασκευάσουμε την κλάση. Ο ελάχιστος αριθμός των διανυσμάτων υποστήριξης παρατηρείται για σ=1.5 και σ=1.75 και είναι 11 και 12, αντίστοιχα.

  Στην πραγματικότητα, μία καλή περιγραφή δεδομένων με ένα μικρό σφάλμα στην κλάση-στόχο απαιτεί λίγα διανύσματα υποστήριξης. Ο καθορισμός της βέλτιστης κλάσης βασίζεται σε ένα συνδυασμό μεταξύ του αριθμού των διανυσμάτων υποστήριξης και του όγκου της υπερσφαίρας. Ο Πίνακας 6.2 δείχνει ότι ο ελάχιστος όγκος σφαίρας παρατηρείται με τη χρήση της $SVDD$ με ελάχιστο αριθμό διανυσμάτων υποστήριξης (12). Αυτή η κλάση είναι η βέλτιστη και χρησιμοποιείται για το σχεδιασμό του K διαγράμματος. Υπολογίζοντας την απόσταση πυρήνων, μπορεί να καθοριστεί το UCL για το K διάγραμμα.



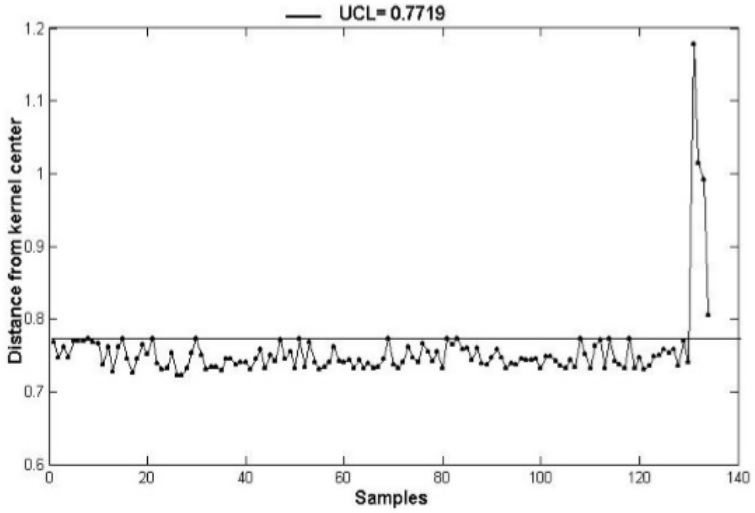
Σχήμα 6.5: Η κατασκευή της βέλτιστης κλάσης χρησιμοποιώντας στον $SVDD$ ταξινομητή.

Πίνακας 6.2: Χαρακτηριστικά της $SVDD$

SVDD	Number of SV	Error on target class	Number of outliers	Sphere volume
SVDD(0.05;1)	22	0.1692	4	2.684
SVDD(0.05;1.25)	18	0.1384	4	2.2586
SVDD(0.05;1.5)	11	0.0846	4	2.0617
SVDD(0.05;1.75)	12	0.0923	4	1.8719

Στη Φάση Ι, η ομάδα δεδομένων χρησιμοποιείται για την κατασκευή της εντός ελέγχου Φάσης για το K διάγραμμα. Στην πραγματικότητα για

f=0.05, έχουμε UCL=0.7719. Αυτό το UCL βασίζεται στα 12 διανύσματα υποστήριξης. Στη Φάση ΙΙ, οι 4 καινούριες εισαγόμενες παρατηρήσεις χρησιμοποιούνται για την ανίχνευση εκτός ελέγχου σημάτων.

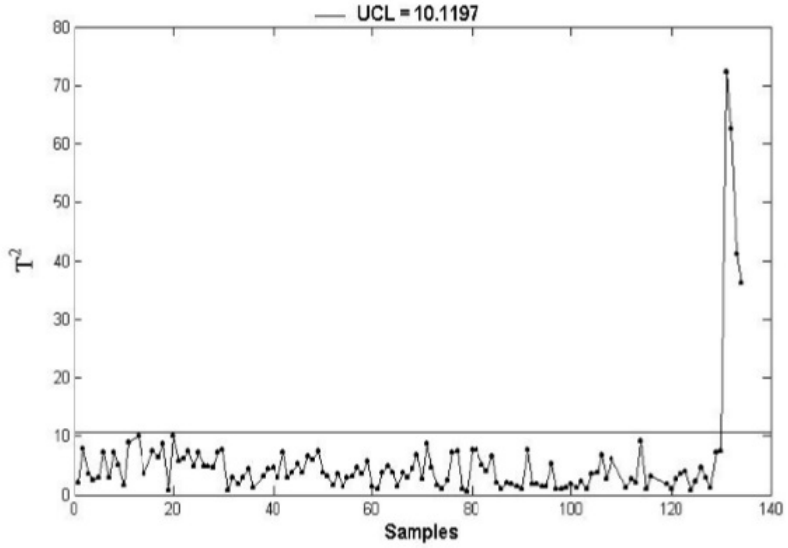


Σχήμα 6.6: Το $K-$διάγραμμα για τη φάση ΙΙ.

Οποιαδήποτε παρατήρηση που έχει απόσταση πυρήνα μεγαλύτερη από την απόσταση από το όριο της κλάσης, θεωρείται εκτός ελέγχου παρατήρηση. Από την άλλη, οι παρατηρήσεις με απόσταση πυρήνα μικρότερη από το όριο θεωρούνται εντός ελέγχου. Στο Σχήμα 6.6 παρουσιάζεται το K-διάγραμμα για τη Φάση ΙΙ, από το οποίο φαίνεται ότι οι παρατηρήσεις 131, 132, 133 και 134 είναι εκτός ελέγχου επειδή ξεπερνούν το όριο. Τέλος στο Σχήμα 6.7 δίνεται το T^2 διάγραμμα ελέγχου για τη Φάση ΙΙ, στο οποίο οι ίδιες παρατηρήσεις είναι πάλι εκτός ελέγχου.

6.3 Εφαρμογή των $rk-$διαγραμμάτων

Το πείραμα των Camci et al. (2008) για την απόδοση των $rk-$διαγραμμάτων, χωρίζεται σε δύο υποενότητες. Πρώτα αξιολογείται το $rk-$διάγραμμα με τη χρήση ομάδων δεδομένων που ακολουθούν κατανομές όπως η κανονική ($normal$), η $lognormal$ και η εκθετική ($exponential$). Στη συνέχεια αξιολογείται το



Σχήμα 6.7: Το διάγραμμα ελέγχου T^2 για τη φάση ΙΙ.

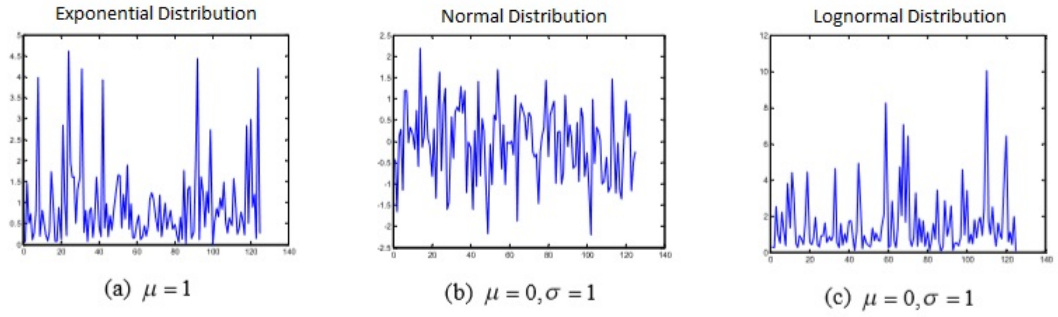
$rk-$διάγραμμα με τη χρήση μίας ομάδας δεδομένων βαθμολόγησης επιδόσεων (όπως τα δεδομένα *Smith*, Smith (1994)) που χρησιμοποιούνται εκτενώς για την ανάπτυξη και την αξιολόγηση στη βιβλιογραφία του ελέγχου διαδικασιών (Chinnam (2002)).
Στον Πίνακα 6.5 και στο Σχήμα 6.8 δίνονται οι παράμετροι των κατανομών και ένα δείγμα των δεδομένων, αντίστοιχα. Για ένα τυπικό $\bar{X}$ διάγραμμα, προτείνεται να σχεδιαστούν τουλάχιστον 20-25 μοτίβα (Woodall and Montgomery (1999)). Εδώ χρησιμοποιούνται 250 δείγματα για κάθε κατανομή, τα οποία ομαδοποιούνται σε ομάδες των 10 και υπολογίζονται ο μέσος και η τυπική απόκλιση για κάθε ομάδα, δίνοντας 25 διδιάστατα σημεία δεδομένων.

Το $rk-$διάγραμμα εξετάζεται με δύο τρόπους. Στον πρώτο μόνο με τα δεδομένα που βρίσκονται εντός ελέγχου και στον δευτέρο με τα εντός ελέγχου και τα οριακά εκτός ελέγχου δεδομένα. Στην δεύτερη περίπτωση, χρησιμοποιούται 10 μη-κανονικά μοτίβα. Τα εκτός ελέγχου δεδομένα δεν είναι απαραίτητα, αλλά βοηθάνε στη βελτίωση της ακρίβειας της μεθόδου. Τα σφάλματα τύπου Ι και ΙΙ καθορίζονται απορρίπτοντας μία σωστή και κάνοντας δεκτή μία λανθασμένη υπόθεση, αντίστοιχα. Στον Πίνακα 6.6 δίνεται ο τρόπος που καθορίζονται αυτά τα δύο σφάλματα.

Όπως φαίνεται από τον Πίνακα 6.7 η μη παραμετρική τεχνική του $rk-$ διαγράμματος μπορεί να ανιχνεύσει τις διεργασίες που είναι εκτός ελέγχου, χωρίς να δίνονται δεδομένα εκτός ελέγχου, με το σφάλμα τύπου Ι να κυμαίνεται από 11.2% μέχρι 26.8% και το σφάλμα τύπου ΙΙ κυμαίνεται από 0% μέχρι 23.9%.

Πίνακας 6.3: Οι παράμετροι των κατανομών για το πείραμα της αξιολόγησης.

		Normal	Lognormal	Exponential
Normal Behavior (Nor)	μ	0	0	1
	σ	1	1	-
Small Mean Shift (SM)	μ	1	1	2
	σ	1	1	-
Large Mean Shift (LM)	μ	3	3	4
	σ	1	1	-
Small Variance Shift (SV)	μ	0	0	-
	σ	2	2	-
Large Variance Shift (LV)	μ	0	0	-
	σ	3	3	-



Σχήμα 6.8: Δείγμα δεδομένων για 3 διαφορετικές κατανομές.

Πίνακας 6.4: Καθορισμός σφαλμάτων τύπου I και II.

Hypothesis Test		**Estimated**	
		In-control	Out-of-control
Actual	*In-control*	Correct	Type-I error
	Out-of-control	Type-II error	Correct

Όταν είναι διαθέσιμα δεδομένα που βρίσκονται εκτός ελέγχου, τα σφάλματα τύπου I και II βελτιώνονται.
Παρουσιάζεται επίσης η αποτελεσματικότητα του $rk-$ διαγράμματος στην ανίχνευση του μέσου και της διασποράς των αλλαγών με τη χρήση της ομάδας

Πίνακας 6.5: Ταξινόμηση ακρίβειας για (Αριστερά) Εκπαίδευση 25 μονοπατιών εντός ελέγχου. (Δεξιά) Εκπαίδευση 25 μονοπατιών εντός ελέγχου και 10 εκτός, όπου το καθένα έχει μία υποομάδα μεγέθους 10.

1) Training only with in-control data

a.1

			Estimated	
			In-control	Out-of-control
Actual	In-control		88.3%	11.7%
	Out-of-control	SM	7.4%	92.6%
		LM	0.0%	100.0%
		SV	6.4%	93.6%
		LV	0.2%	99.8%

2) Training with in-control and limited out-of-control data

a.2

			Estimated	
			In-control	Out-of-control
Actual	In-control		86.6%	13.4%
	Out-of-control	SM	9.0%	91.0%
		LM	0.0%	100.0%
		SV	4.1%	95.9%
		LV	0.1%	99.9%

a) Normal Distribution

b.1

			Estimated	
			In-control	Out-of-control
Actual	In-control		73.2%	26.8%
	Out-of-control	SM	4.5%	95.5%
		LM	0.0%	100.0%
		SV	14.8%	85.2%
		LV	4.1%	95.9%

b.2

			Estimated	
			In-control	Out-of-control
Actual	In-control		81.3%	18.7%
	Out-of-control	SM	9.5%	90.5%
		LM	0.0%	100.0%
		SV	16.9%	83.1%
		LV	3.0%	97.0%

b) Lognormal Distribution

c.1

			Estimated	
			In-control	Out-of-control
Actual	In-control		88.8%	11.2%
	Out-of-control	SM	23.9%	76.1%
		LM	0.7%	99.3%

c.2

			Estimated	
			In-control	Out-of-control
Actual	In-control		80.8%	19.2%
	Out-of-control	SM	16.0%	84.0%
		LM	0.4%	99.6%

c) Exponential Distribution

δεδομένων βαθμολόγησης επιδόσεων που προτείνται από τον Smith (1994) και συγκρίνονται τα δεδομένα με μία γενική SVM μέθοδο, η οποία προτείνεται από τον Chinnam (2002) και ένα MLP μοντέλο νευρωνικού δικτύου (Smith (1994)). Η ομάδα δεδομένων έχει 300 δείγματα από την εντός και την εκτός ελέγχου κατάσταση με μεγάλες και μικρές αλλαγές μέσου (LM και SM, αντίστοιχα), μεγάλες και μικρές αλλαγές διασποράς (LV και SV, αντίστοιχα). Οι παράμετροι δίνονται στον Πίνακα 6.8.

Τα αποτελέσματα χωρίζονται σε 3 διαφορετικές κατηγορίες. Στην πρώτη κατηγορία, τα αποτελέσματα του $rk-$διαγράμματος συγκρίνονται με τα αποτε-

Πίνακας 6.6: Παράμετροι της ομάδας δεδομένων που εξετάζεται.

State Label	μ	σ
Normal Behavior (Nor)	0	1
Small Mean Shift (SM)	1	1
Large Mean Shift (LM)	3	1
Small Variance Shift (SV)	0	2
Large Variance Shift (LV)	0	3

λέσματα του *Shewart* διαγράμματος, του MLP και του SVM. Στη δεύτερη κατηγορία, αναφέρονται τα αποτελέσματα για δύο διαφορετικά $rk-$διαγράμματα, το $rk-$διάγραμμα, στο οποίο εκπαιδεύονται μόνο εντός ελέγχου δεδομένα και το $rk-$διάγραμμα, στο οποίο εκπαιδεύονται παράλληλα και εκτός ελέγχου δείγματα. Στην τρίτη κατηγορία, αναφέρονται τα αποτελέσματα του $rk-$διαγράμματος που εκπαιδεύεται με πολύ περιορισμένα εντός και εκτός ελέγχου δεδομένα.

Στην πρώτη κατηγορία, η ακρίβεια της ταξινόμησης του $rk-$διαγράμματος

Πίνακας 6.7: Η ταξινόμηση της ακρίβειας για το $rk-$διάγραμμα, το MLP, το *Shewhart* και το SVM διάγραμμα με τη χρήση των δεδομένων βαθμολόγησης επιδόσεων που προτείνει ο *Smith*.

	rk-Chart		MLP	Shewhart Control Charts	SVM	
	Test	Train	Test	Test	Test	Train
Small Shift	91%	92%	72%	73%	93%	91%
Large Shift	96%	N/A	100%	100%	100%	99%

υπολογίζεται για να συγκριθούν τα αποτελέσματα των διάφορων μεθόδων. Όπως φαίνεται από τον Πίνακα 6.9, το $rk-$διάγραμμα αποδίδει καλύτερα από τα διαγράμματα MLP και *Shewhart*. Το SVM είναι μόνο 1% με 4% καλύτερο από το $rk-$διάγραμμα. Παρά το γεγονός ότι η SVM μέθοδος δίνει καλύτερο αποτέλεσμα σε σχέση με το $rk-$διάγραμμα, υπάχουν δύο δυσκολίες στην παρουσίαση του SVM και του MLP στις εφαρμογές του πραγματικού κόσμου.

1. Αναπτύσσεται ένα ξεχωρίστό μοντέλο για κάθε εκτός ελέγχου κατάσταση του SVM και του MLP. Υπάρχουν 4 εκτός ελέγχου καταστάσεις (SM, LM, SV, LV) που έχουν σαν αποτελέσμα τα 4 SVM και MLP

μοντέλα. Παρά το γεγονός ότι κάθε μοντέλο δουλεύει για την ανάπτυξη των εκτός ελέγχου καταστάσεων, δεν μπορούν να δουλέψουν αποτελεσματικά για άλλες εκτός ελέγχου καταστάσεις που υπάρχει ήδη στο μοντέλο. Επιπρόσθετα, τα SVM και MLP είναι πιο ευαίσθητα σε μία εκτός ελέγχου κατάσταση που δεν έχει καθοριστεί. Αντίθετα το $rk-$διάγραμμα χαρακτηρίζει την εντός ελέγχου κατάσταση της διεργασίας και έχει την ικανότητα να ανιχνεύει κάθε μη καθορισμένο τύπο εκτός ελέγχου κατάστασης.

2. Το SVM και το MLP εκπαιδεύονται με 300 δείγματα από μία εντός ελέγχου κατάσταση και άλλα 300 από κάθε εκτός ελέγχου κατάσταση, οπότε συνολικά χρησιμοποιούνται 2400 δείγματα. Από την άλλη, το $rk-$διάγραμμα χρησιμοποιεί μόνο 300 δείγματα από την εντός και 200 από τις εκτός ελέγχου καταστάσεις.

Πίνακας 6.8: Τα σφάλματα τύπου I και II για τα δεδομένα βαθμολόγησης επιδόσεων που προτείνει ο $Smith$ χρησιμοποιώντας την εντός ελέγχου κατάσταση.

			Estimated							
			rk-Chart$^{(1)}$				rk-Chart$^{(2)}$			
			In*	Out*	In*	Out*	In*	Out*	In*	Out*
			Train		Test		Train		Test	
Actual	In-cont.		100%	0%	84%	16%	88%	12%	90%	10%
	Out*	SM	N/A	N/A	5%	95%	1%	99%	7%	93%
		LM	N/A	N/A	0%	100%	N/A	N/A	0%	100%
		SV	N/A	N/A	15%	85%	4%	96%	9%	91%
		LV	N/A	N/A	15%	85%	N/A	N/A	8%	92%

Στην δεύτερη κατηγορία, αναφέρονται τα αποτελέσματα του $rk-$διαγράμματος χωρίς εκτός ελέγχου δεδομένα. Στην πρώτη περίπτωση ($rk - Chart^{(1)}$), εκπαιδεύτηκαν μόνο 300 δείγματα από τα εντός ελέγχου δεδομένα, ενώ στην δεύτερη περίπτωση ($rk - Chart^{(2)}$) δίνονται για εκπαίδευση 300 δείγματα από εντός ελέγχου δεδομένα και 100 δείγματα από μικρές αλλαγές στον μέσο και 100 δείγματα από μικρές αλλαγές στη διασπορά. Τα σφάλματα τύπου I και II που δίνονται στον Πίνακα 6.8 υπόσχονται σφάλμα τύπου I στο 16% και τύπου II το πολύ 15% στο $rk - Chart^{(1)}$ και σφάλμα τύπου I στο 12% και τύπου II το πολύ 9% με το $rk - Chart^{(2)}$.
Στην τρίτη κατηγορία, το $rk-$διάγραμμα παρουσιάζεται με πολύ περιορισμένα εντός ελέγχου δεδομένα. Παρουσιάζεται με 25 μοτίβα από τα δεδομένα βαθμολόγησης επιδόσεων και τα αποτελέσματα δίνονται στον Πίνακα 6.9.

Όπως φαίνεται από τον Πίνακα 6.9, στην χειρότερη από τις περιπτώσεις που

Πίνακας 6.9: Τα στάδια διαδικασίας ταξινόμησης για μη συσχετισμένα δεδομένα *Smith* με περιορισμένο αριθμό για εντός και εκτός ελέγχου δείγματα.

Data size of 25			*rk*-Chart	
			in*	out*
Actual	in		88%	12%
	out	SM	9%	91%
		LM	1%	99%
		SV	8%	92%
		LV	4%	96%

ελέγχονται, το $rk-$διάγραμμα μπορεί να ανιχνεύει αποτελεσματικά τις εκτός ελέγχου καταστάσεις (με σφάλμα τύπου II 9%), με τις λάθος προειδοποιήσεις να είναι μόνο 12% ακόμα και με περιορισμένο μέγεθος δεδομένων.
Σχετικά με την επιλογή των παραμέτρων, όταν το διάγραμμα ελέγχου αρχικοποείται με εντός ελέγχου δεδομένα, απαιτούνται δύο παράμετροι η τιμή ποινής για τη λανθασμένη ταξινόμηση των εντός ελέγχου δεδομένων (C) και η συμπάγεια (ς). Οι προτεινόμενες τιμές είναι C=1.0 και ς=0.2. Προτείνεται μία πρόσθετη παράμετρος ποινής (C_0) για τις περιπτώσεις εκτός ελέγχου δεδομένων για αρχικοποίηση ($C_0 = 1.0$). Όλα τα πειραματικά αποτελέσματα που αναφέρονται βασίζονται σε αυτές τις ρυθμίσεις.

6.4 Σύγκριση μεταξύ του K και του T^2 διαγράμματος ελέγχου.

Αυτή η σύγκριση γίνεται βάσει του κριτηρίου ARL. Η τιμή του ARL υπολογίζεται με τεχνική προσομοίωσης. Με αυτό τον τρόπο, δημιουργούνται 4 πολυμεταβλητές κανονικές μεταβλητές. Κάθε μεταβλητή αποτελείται από 10000 παρατηρήσεις. Για τη μελέτη της ευαισθησίας των K και T^2 διαγραμμάτων, προτείνονται 3 τύποι αλλαγής του μέσου: μικρές, μεσαίες και μεγάλες μετατοπίσεις.

Ο Πίνακας 6.12 δείχνει ότι το K-διάγραμμα είναι πιο ευαίσθητο από το T^2 για τις μικρές και τις μεσαίες αλλαγές του διανύσματος του μέσου. Πράγματι, η διαφορά του επιπέδου ευαισθησίας μεταξύ αυτών των διαγραμμάτων μπορεί να εξηγηθεί από τη φύση της απόστασης που δίνεται από τις δύο μεθόδους. Ο πυρήνας απόστασης που χρησιμοποιείται με το $K-$διάγραμμα, δημιουργεί μία

Πίνακας 6.10: Το ARL για το K και το T^2 διάγραμμα ελέγχου.

Shifts	ARL for k-chart	ARL for T^2 chart
Small	5.8858	8.4388
Medium	3.3069	4.8591
Large	1	1

συμπαγή απόσταση σε σύγκριση με την απόσταση που χρησιμοποιείται με το T^2 διάγραμμα. Ένας άλλος παράγοντας που μπορεί να εξηγεί την ευαισθησία του $K-$διαγράμματος είναι η φύση του ορίου ελέγχου που βασίζεται στα διανύσματα υποστήριξης, τα οποία είναι πολύ ευαίσθητα σε οποιαδήποτε αλλαγή στο πλάτος του πυρήνα.

6.5 Μελέτη προσομοίωσης των D^2, K^2, T^2 και $OC - SVM$

Παρακάτω γίνεται σύγκριση της απόδοσης των D^2, T^2, K^2 και $OC - SVM$ διαγραμμάτων Sukchotrat et al. (2009). Τα δεδομένα βασίζονται σε δείγμα δεδομένων διδιάσταστης κανονικής, διδιάστατης t, διδιάστατης $gamma$ και $banana-shaped$ κατανομής. Για τα D^2 και τα $OC-SVM$ διαγράμματα, χρησιμοποιείται το πλάτος του $Gaussian$ πυρήνα, $S = 1$ για τις περιπτώσεις της κανονικής, της $gamma$ και της t κατανομής και $S = 3$ για τα $banana-shaped$ δεδομένα. Για τα K^2 διαγράμματα, χρησιμοποιείται $k = 30$.
Παρακολουθούνται 1000 παρατηρήσεις της Φάσης ΙΙ (900 εντός ελέγχου και 100 εκτός παρατηρήσεις) βάσει των ορίων ελέγχου που έχουν σταθεροποιηθεί από 200 παρατηρήσεις της Φάσης Ι. Έστω ότι μ_0 και Σ_0 είναι το διάνυσμα του μέσου και ο πίνακας συνδιασποράς για τα εντός ελέγχου δεδομένα, αντίστοιχα, και ότι $\mu_1 = \mu_0 + \delta$ είναι το διάνυσμα του μέσου για τα εκτός ελέγχου δεδομένα. Το μέγεθος της μετατόπισης δ δίνεται από την παράμετρο $\lambda = \sqrt{\delta^T \Sigma_0^{-1} \delta}$.
Για την παραγωγή των εκτός ελέγχου δεδομένων για τις κατανομές που αναφέρθηκαν, θεωρούνται δύο τύποι μετατόπισης των μέσων ($\lambda = 2$ και $\lambda = 3$). Για μία συγκεκριμένη τιμή του λ όλες οι μεταβλητές αλλάζουν το ίδιο. Παρακάτω περιγράφονται τα αποτελέσματα των σεναρίων προσομοίωσης

1. N_2, $\lambda = 2$. Η περίπτωση μέτριας μετατόπισης του μέσου της διδιάστατης κανονικής κατανομής με $\mu_0 = \begin{bmatrix} 0 & 0 \end{bmatrix}$ και $\Sigma_0 = \begin{bmatrix} 1 & 0.35 \\ 0.35 & 1 \end{bmatrix}$.

2. N_2, $\lambda = 3$. Η περίπτωση μεγάλης μετατόπισης του μέσου της διδιάστατης

κανονικής κατανομής με $\mu_0 = \begin{bmatrix} 0 & 0 \end{bmatrix}$ και $\Sigma_0 = \begin{bmatrix} 1 & 0.35 \\ 0.35 & 1 \end{bmatrix}$.

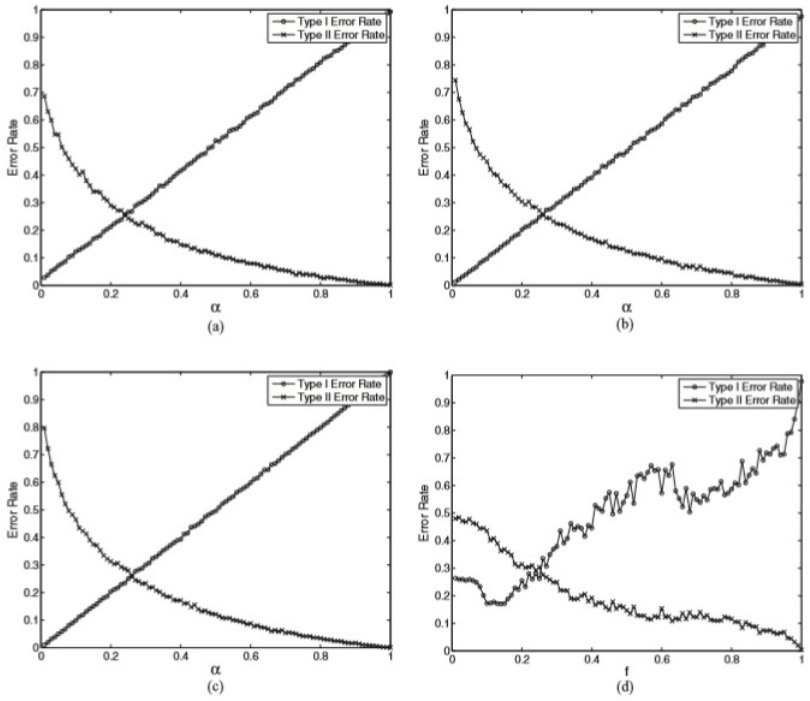
3. $t_2(3)$, $\lambda = 2$. Η περίπτωση μέτριας μετατόπισης του μέσου της διδιάστατης t κατανομής με τρεις βαθμούς ελευθερίας.

4. $t_2(3)$, $\lambda = 2$. Η περίπτωση μεγάλης μετατόπισης του μέσου της διδιάστατης t κατανομής με τρεις βαθμούς ελευθερίας.

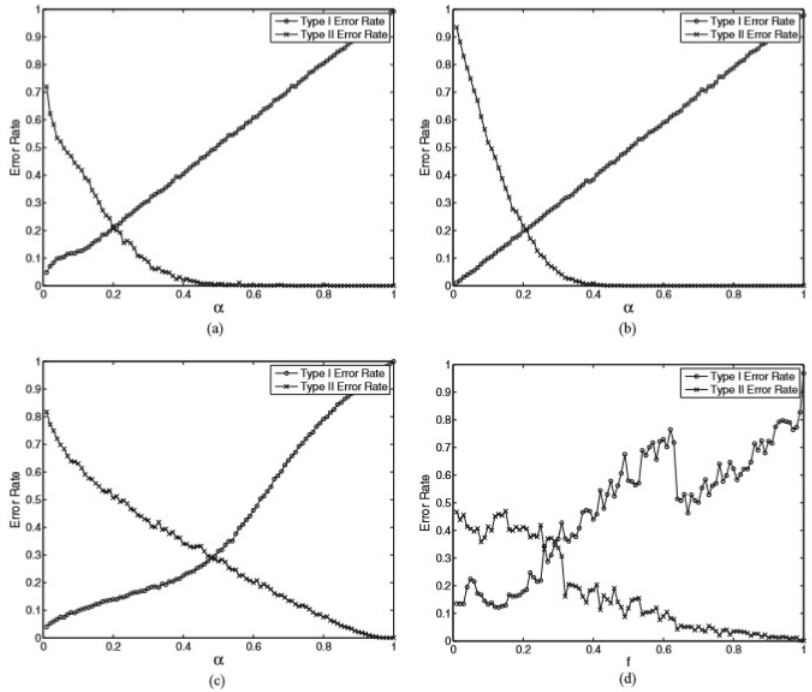
5. $Gam_2(1,1)$, $\lambda = 2$. Η περίπτωση μέτριας μετατόπισης του μέσου της διδιάστατης *gamma* με τις παραμέτρους *scale* και *shape* να είναι ίσες με τη μονάδα.

6. $Gam_2(1,1)$, $\lambda = 3$. Η περίπτωση μεγάλης μετατόπισης του μέσου της διδιάστατης *gamma* με τις παραμέτρους *scale* και *shape* να είναι ίσες με τη μονάδα.

7. $Banana - shaped$. Μία ομάδα $banana - shaped$ δεδομένων με δύο διαφορετικές γωνίες.



Σχήμα 6.9: Η μέση τιμή του σφάλματος τύπου I και II για τα (a) D^2, (b) K^2, (c) T^2 και (d) $OC - SVM$ διαγράμματα (N_2 με λ=2).



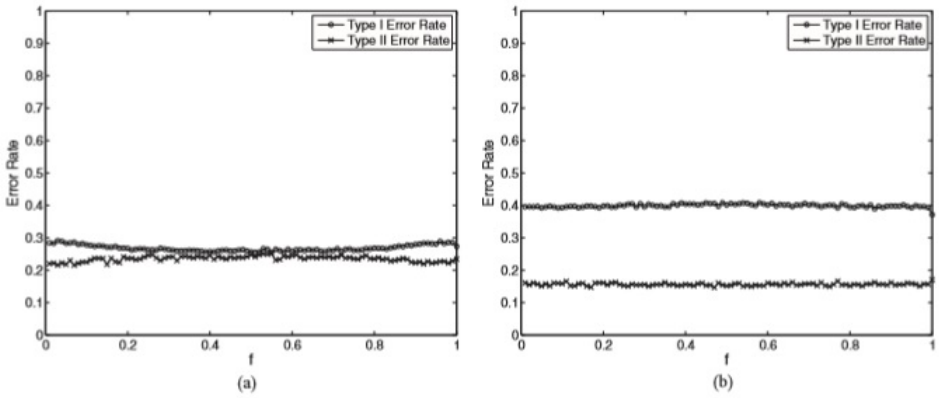
Σχήμα 6.10: Η μέση τιμή του σφάλματος τύπου Ι και ΙΙ για τα (a) D^2, (b) K^2, (c) T^2 και (d) $OC-SVM$ διαγράμματα (Gam_2(1,1) με λ=2).

6.5.1 Τα όρια ελέγχου.

Σε αντίθεση με τα υπάρχοντα $OC-SVM$ διαγράμματα που χρησιμοποιούν την παράμετρο f για να προσθέσουν τα όρια ελέγχου, τα όρια ελέγχου των διαγραμμάτων D^2 και K^2 προστίθονται από το ποσοστό, το οποίο εκτιμάται από τη μέθοδο *bootstrap*. Τα Σχήματα 6.9 και 6.10 δείχνουν πως ελέγχονται τα σφάλματα τύπου Ι και ΙΙ για τα διαγράμματα D^2, K^2, T^2 και $OC-SVM$, από τους παράγοντες ελέγχου (α ή f). Χρησιμοποιούμε τις μέσες τιμές των πραγματικών σφαλμάτων τύπου Ι και ΙΙ από 100 προσομοιώσεις. Το τυπικά σφάλματα των 100 προσομοιώσεων είναι σχετικά μικρά (μεταξύ 0.02 και 0.06), αποδεικνύοντας ότι οι 100 προσομοιώσεις είναι αρκετές για να βγει ένα ουσιαστικό συμπέρασμα. Παρουσιάζονται τα αποτελέσματα μόνο για τα N_2 και $Gam_2(1,1)$ σενάρια, αντίστοιχα, σαν παραδείγματα κανονικών και μη περιπτώσεων. Γενικά, όσο αυξάνεται ο παράγοντας ελέγχου, όλα τα διαγράμματα ελέγχου παράγουν μεγαλύτερο σφάλμα τύπου Ι αλλά μικρότερο σφάλμα τύπου ΙΙ. Η ιδιαίτερα ισχυρή θετική συσχέτιση μεταξύ της πραγματικής τιμής του σφάλματος τύπου Ι και του παράγοντα ελέγχου είναι επιθυμητή. Τα διαγράμματα D^2 και K^2 ικανοποιούν αυτή την προυπόθεση και στην κανονική αλλά και στην μη κα-

νονική περίπτωση, αλλά το T^2 διάγραμμα μόνο στην κανονική. Το διάγραμμα $OC-SVM$ αποτυγχάνει να δείξει υψηλή γραμμική συσχέτιση μεταξύ της πραγματικής τιμής του σφάλματος I και του παράγοντα ελέγχου. Επιπλέον, οι τιμές των σφαλμάτων τύπου I και II μπορεί να μην ελέγχονται κατάλληλα από το f όσο το μέγεθος του στόχου των παρατηρήσεων αυξάνεται στα $OC-SVM$ διαγράμματα.

Το Σχήμα 6.11 δείχνει το $OC-SVM$ διάγραμμα που έχει κατασκευαστεί από



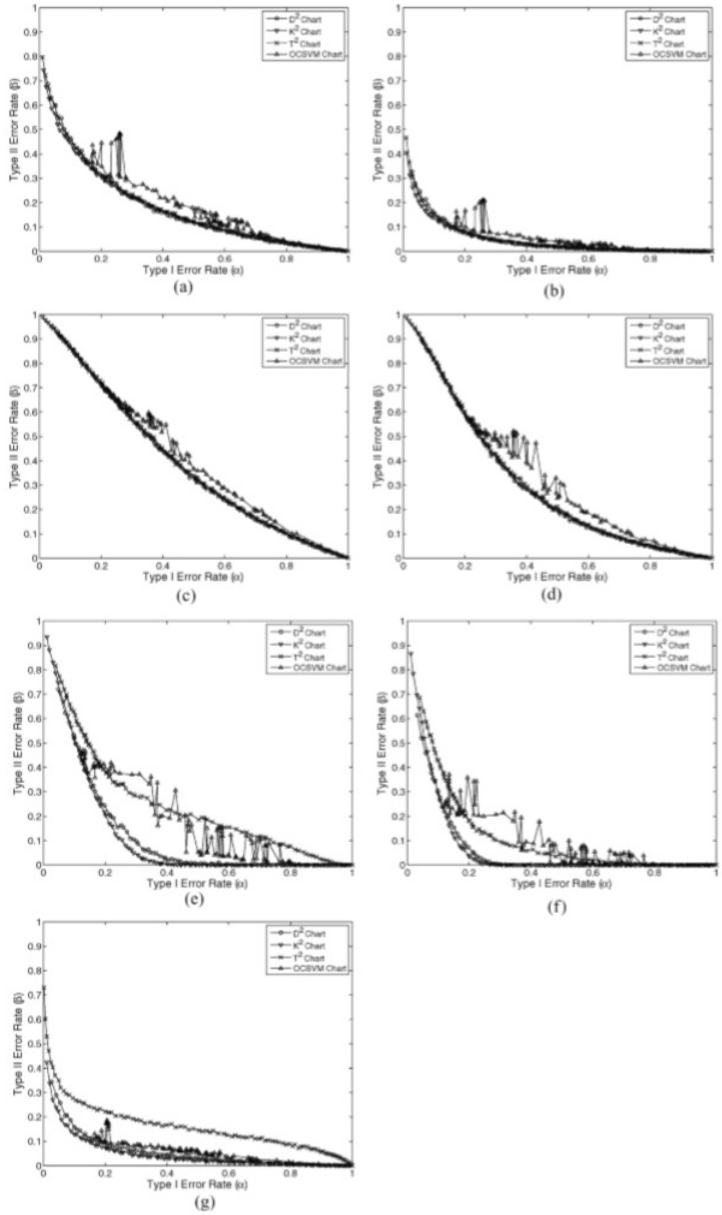
Σχήμα 6.11: Η μέση τιμή του σφάλματος τύπου I και II για τα διαγράμματα $OC-SVM$, όταν ο αριθμός των παρατηρήσεων στη φάση I είναι μεγάλος (N_2 με λ=2) (a) 300 παρατηρήσεις και (b) 400 παρατηρήσεις.

την N_2 με $\lambda = 2$ και με 300 και 400 παρατηρήσεις ως στόχο. Παρατηρείται ότι τα σφάλματα τύπου I και II φαίνονται να είναι σταθερά για τις διάφορες τιμές του f. Με μεγάλους αριθμούς στόχου παρατηρήσεων, η τιμή του f παίζει μικρό ρόλο στην αλλαγή των ορίων ελέγχου και οδηγεί σε σχετικά σταθερές τιμές για τα σφάλματα τύπου I και II. Το αποτέλεσμα είναι ότι η f είναι ακατάλληλη επιλογή για τον παράγοντα ελέγχου στα $OC-SVM$ διαγράμματα.

6.5.2 Συγκρίσεις των επιδόσεων.

Παρακάτω συγκρίνονται οι μέσες τιμές των σφαλμάτων τύπου I και II για τις 100 προσομοιώσεις για τα διαγράμματα D^2, K^2, T^2 και $OC-SVM$. Το διάγραμμα ελέγχου που δίνει τη μικρότερη τιμή του σφάλματος τύπου II θεωρείται σαν καλύτερη μέθοδος αν η τιμή του σφάλματος τύπου II είναι παρόμοια. Το Σχήμα 6.12 παρουσιάζει τις μέσες τιμές των σφαλμάτων τύπου I και II για τα

σενάρια προσομοίωσης που αναφέρθηκαν.



Σχήμα 6.12: Τα σφάλματα τύπου I και II για τα διαγράμματα D^2, K^2, T^2 και $OC-SVM$ για τα σενάρια προσομοίωσης που αναφέρθηκαν. (a) N_2 με λ=2, (b) N_2 με λ=3, (c) t_2 με λ=2, (d) t_2 με λ=3, (e) Gam_2(1,1) με λ=2, (f) Gam_2(1,1) με λ=3 και (g) $banana-shaped$.

Τα αποτελέσματα δείχνουν ότι τα διαγράμματα D^2 και K^2 παράγουν μικρότερο σφάλμα τύπου II σε σχέση με το T^2 διάγραμμα, ενώ δίνουν παρόμοια τιμή σφάλματος τύπου I στα $gamma$ και $banana-shaped$ δεδομένα. Στην περίπτωση της κανονικής και της t κατανομής, όλες οι μέθοδοι έχουν συγκρίσιμες αποδόσεις. Το έυρος των τυπικών σφαλμάτων για τις 100 προσομοιώσεις είναι μεταξύ του 0.02 και του 0.08 για της περίπτωση της κανονικής, της t και της $gamma$ κατανομής, ενώ στην περίπτωση του διαγράμματος $OC - SVM$ παρατηρούνται μεγαλύτερα τυπικά σφάλματα (μεταξύ 0.1 και 0.26). Πρέπει να σημειωθεί ότι τα $OC - SVM$ διαγράμματα παράγουν ακανόνιστες τιμές σφάλματος τύπου II για τις τιμές σφάλματος τύπου I. Συνεπώς είναι δύσκολη η σύγκριση της απόδοσης του $OC - SVM$ με τα άλλα διαγράμματα.

Βιβλιογραφία

[1] Κουκουβίνος, Χ. (2016). *"Στατιστικός Έλεγχος Ποιότητας"*. ΕΜΠ

[2] Simon Haykin, *Νευρωνικά Δίκτυα και Μηχανική Μάθηση* (2010), Παπασωτηρίου, 2010.

[3] Alt, F. B. (1985). *Multivariate Quality Control*, *in Encyclopedia of Statistical Sciences*, Vol. 6, N. L. Johnson and S. Kotz, (eds.) Wiley, New York.

[4] F. Aparisi, G. Avendano, and J. AndSanz, Techniques to interpret T2 control chart signals, *IIE Transactions*, vol. 38, no. 8, pp. 647-657, 2006

[5] F. Aparisi, "Hotelling's T^2 Control chart with adaptive sample size," *International Journal of Production Research*, vol. 34, no. 10, pp. 2853-2862, Oct 1996.

[6] F. Aparisi and C.L. Haro, "Hotelling's T^2 Control chart with sampling intervals", *International Journal of Production Research*, vol. 39, no. 14, pp. 3127-3140, 2001.

[7] F. Aparisi and C.L. Haro, "A comparison of T^2 Control chart with variable sampling schemes as opposed to MEWMA chart," *International Journal of Production Research*, vol. 41, no. 10, pp. 2169-2182, 2003.

[8] Aradhye, H. B., Bakshi, B. R., Strauss, R. A. and Davis, J. F. (2001). *Multiscale statistical process control using wavelets: Theoretical analysis and properties*. Columbus, OH, Ohio State University.

[9] B. E. Boser and I. M. Guyon and V. N. Vapnik, "A Training Algorithm for Optimal Margin Classifiers". Proceedings of the 5th annual ACM workshop on computational learning theory (1992), pp. 144-152, ACM Press

[10] Breunig, M.M., Kriegel, H.P., Ng, R.T. and Sander, J. (2000) LOF: identifying density-based local outliers. *In Proceedings of the ACM SIGMOD 2000 international conference on management of data*, 29, pp. 93–104.

[11] F. Camci, R.B. Chinnam and R.D. Ellis, "Robust kernel distance multivariate control chart using support vector principles," *International Journal of Production Research*, vol. 46, no. 18, pp. 5075-5095, 2008.

[12] Chakraborti, S., Van der Laan, P. and Van de Wiel, M., A class of distribution-free control charts. *Journal of the Royal Statistical Society*, Series C, 2004, 55(3), 443-462.

[13] Chakraborti, S., Van der Laan, P. and Bakir, S. T., Nonparametric control charts: An overview and some results. *Journal of Quality Technology*, 2001, 33, 304-315.

[14] Y.K. Chen and K.C. Chiou, "Adaptive sampling enhancement for Hotelling's T^2 charts," *The 2005 International Conference in Management Sciences and Decision Marking*, Tamkang University, pp. 281-296, 2005.

[15] Cheng, C. S., A neural network approach for the analysis of control chart patterns. *International Journal of Production Research* 35 (1997) 667-697.

[16] Cheng Zhi-Qiang, Ma Yi-Zhong, Bu Jing (2010). Mean Shifts Identification Model in Bivariate Process Based on LS-SVM Pattern Recognizer, *International Journal of Digital Content Technology and its Applications*. Volume 4, Number 3.

[17] Cheng H.P, Cheng C.S., A Support Vector Machine for Recognizing Control Chart Patterns in Multivariate Processes, in *Proceedings of the 5th Asian Quality Congress*, Incgeon, Korea, pp 456-465, 2007.

[18] C.S. Cheng and H.P. Cheng, "Identifying the source of variance shifts in the multivariate process using neural networks and support vector machines," *Expert Systems with Applications*, vol. 35, no.1-2, pp. 198-206, Jul-Aug 2008.

[19] Chinnam, R. B., "Support vector machines for recognizing shifts in correlated and other manufacturing processes". *International Journal of Production Research*, 2002, 40, 4449 - 4466.

[20] Chinnam, R. B. and Kolarik, W. J., Automation and the total quality paradigm, in *Proceedings of the 1st IERC*, 1992, Chicago, IL, IIE.

[21] Cook, D. F. and Chiu, C., Using radial basis function neural networks to recognize shifts in correlated manufacturing process parameters. *IIE Transactions*, 1998, 30, 227-234.

[22] Cristianini, N. and Taylor, J. S., *An introduction to support vector machines and other kernel-based learning methods*. (Cambridge: Cambridge University Press).

[23] Walid Gani , Hassen Taleb , Mohamed Limam. "Statistical process control using support vector machines: A case study", *International Journal of Quality and Standards*, 2(1), pp. 122-138 , 2008.

[24] Ge, Z. and Song, Z. (2008). Online monitoring of nonlinear multiple mode processes based on adaptive local model approach. *Control Engineering Practice*, 16, 1427–1437.

[25] Ge, Z., Song, Z. (2007). Process monitoring based on independent component analysis−principal component analysis (ICA−PCA) and similarity factors. *Industrial and Engineering Chemistry Research*, 46, 2054-2063.

[26] Guh, R. S., Hsieh, Y, C., A neural network based model for abnormal pattern recognition of control charts. *Computers & Industrial Engineering*, 36 (1999) 97-108.

[27] Trevor Hastie, Robert Tibshirani, Jerome Friedman, 2009. *The Elements of Statistical Learning : Data Mining, Inference and Prediction*. 2^{nd} edition. Springer Series in Statistics.

[28] Hawkins, D. M. (1981). A CUSUM for a Scale Parameter, *Journal of Quality Technology*, Vol. 13(4), pp. 228–235.

[29] Hawkins, D. M. (1993). Cumulative Sum Control Charting: An Underutilized SPC Tool, *Quality Engineering*, Vol. 5(3), pp. 463–477.

[30] Holmes, D. S., and A. E. Mergen (1993). Improving the Performance of the T^2 Control Chart, *Quality Engineering*, Vol. 5(4), pp. 619–625.

[31] Chun-Chin Hsu, Mu-Chen Chen, Long-Sheng Chen. Integrating independent component analysis and support vector machine for multivariate process monitoring. *Computers and Industrial Engineering*, 59 (2010) 145–156.

[32] M. Kano, S. Tanaka, S. Hasebe, I. Hashimoto, and H. Ohno, "Monitoring independent components for fault detection," *AIChE Journal*, vol. 49, no. 4, pp. 969–976, 2003.

[33] Murray, W., Gill, P. and Wright, M. (1981). *Practical Optimization*, Academic Press.

[34] Kuhn, H. and Tucker, A., Nonlinear programming, *in Proceedings of 2nd Berkeley Symposium on Mathematical Statistics and Probabilistics*, 1951, Berkeley, CA, University of California Press, 481-492.

[35] S. Kumar, A.K. Choudhary, M. Kumar, R. Shankar and M.K. Tiwari, "Kernel distance-based robust support vector methods and its application in developing a robust K-chart," *International Journal of Production Research*, vol. 44, no. 1,pp. 77-96, 2006.

[36] Lee, J. M., Qin, S. J. and Lee, I. B. (2006). Fault detection and diagnosis based on modified independent component analysis. *AIChE Journal*, 52(10), 3501–3514.

[37] J. M. Lee, C. Yoo, and I. B. Lee, "Statistical process monitoring with independent component analysis," *Journal of Process Control*, vol. 14, no. 5, pp. 467–485, 2004.

[38] Liu, R. Y. and Singh, K., A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association* (JASA), 1993, 88, 252-260.

[39] Chia-Hau Liu, Tai-Yue Wang, An AK-chart for the Non-Normal Data,World Academy of Science, Engineering and Technology International *Journal of Computer, Electrical, Automation, Control and Information Engineering*, Vol:8, No:7, 2014.

[40] Lowry, C. A., W. H. Woodall, C. W. Champ, and S. E. Rigdon (1992). A Multivariate Exonentially Weighted Moning Average Control Chart, Technometrics, Vol. 34(1), pp. 46–53.

[41] Mangasarian, O.L., Non-linear Programming, 1994 (Society for Industrial and Applied Mathematics: Philadelphia, PA).

[42] R. L. Mason, N. D. Tracy, and J. C. Young, Decomposition of T^2 for multivariate control chart interpretation, *Journal of Quality Technology*, vol. 27, pp. 99–105, 1995.

[43] Mercer, J., Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London*, Series A, 1909, 209, 415-446.

[44] Messaoud, A., Weihs, C. and Hering, F. (2004). *A nonparametric multivariate control chart based on data depth*. Dortmund, Germany, Department of Statistics, University of Dortmund.

[45] Montgomery, D.C., 2013, *Introduction to Statistical Quality Control*, John Wiley, New York, 7th edition.

[46] Müller, K. R., Mika, S., Rätsch, G., Tsuda, K. and Schölkopf, B., An introduction to kernel-based learning algorithms. *IEEE Neural Networks*, 2001, 12, 181-201.

[47] S. T. A. Niaki and B. Abbasi, "Fault diagnosis in multivariate control charts using artificial neural networks," *Quality and Reliability Engineering International*, vol. 21, no. 8, pp. 825-840, 2005.

[48] X. Ning and F. Tsung, "Improved design of kernel distance-based charts using support vector method," *IIE Transactions*, vol. 45, no. 4, pp. 467-476, Apr 2013.

[49] Page, E. S. (1954). Continuous Inspection Schemes, *Biometrics* ,Vol. 41(1), pp. 100–115.

[50] Pignatiello, J., Runger, G.: Comparisons of multivariate CUSUM charts. *J. Qual. Tech.* 22(3), 173–186 (1990)

[51] Polansky, A.M., A smooth nonparametric approach to multivariate process capability. *Technometrics*, 2001, 43, 199–211

[52] Pugh, G. A., A comparison of neural networks to spc charts. *Computers and Industrial Engineering*, 1991, 21, 253-255.

[53] P. Qiu, (2014). Introduction to Statistical Process Control. Boca Raton, FL, CRC Press.

[54] P. Qiu and D. Hawkins, "A rank based multivariate CUSUM procedure," *Technometrics*, vol. 43, no. 2, pp. 120-132,May 2001.

[55] P. Qiu and D. Hawkins, "A nonparametric multivariate CUSUM procedure for detecting shifts in all directions," *Journal of the Royal Statistical Society-D*, vol. 52,pp. 151-164, 2003.

[56] Roberts, S. W. (1959). Control Chart Tests Based on Geometric Moving Averages, *Technometrics*, Vol. 42(1), pp. 97–102.

[57] Rose, K., Mathematics of success and failure. Circuits and Devices, IEEE, 1991, 7, 26-30.

[58] Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths, *Annals of Statistics* 35(3): 1012–1030.

[59] Rosset, S., Zhu, J. and Hastie, T. (2004). Margin maximizing loss functions, in S. Thrun, L. Saul and B. Schoelkopf (eds), *Advances in Neural Information Processing Systems 16*, MIT Press, Cambridge, MA.

[60] Edgar Santos-Fernandez (2012), *Multivariate Statistical Quality Control Using R*, New York Springer.

[61] Schilling, E. G. and Nelson, P.R., The effect of non-normality on the control limits of X charts. *Journal of Quality Technology*, 1976, 8, 183-187.

[62] Schoelkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J. and Williamson, R.C. (2001) Estimating the support of a high-dimensional distribution. *Neural Computation*, 13, 1443–1471.

[63] Yuehjen E. Shao. Using a Computational Intelligence Hybrid Approach to recognize the Faults of Variance Shifts for a Manufacturing Process. *Journal of Industrial and Intelligent Information*, Vol. 4, No. 2, March 2016.

[64] Y . E. Shao and B. S. Hsu, Determining the contributors for a multivariate SPC chart signal using artificial neural networks and support vector machine, International *Journal of Innovative Computing, Information and Control*, vol. 5, pp. 4899–4906, 2009.

[65] Slišković, Dražen, Grbić, Ratko, Hocenski, Željko. Multivariate statistical process monitoring. *Tehnicki Vjesnik-Technical Gazette*, 19 (2012) , 1, 33-41.

[66] Smith, A. E., "X-bar and r control chart interpretation using neural computing". *International Journal of Production Research*, 1994, 32, 309-320.

[67] Stoumbos, Z. G. and Reynolds, M. R., On shewhart-type nonparametric multivariate control charts based on data depth. *Frontiers in Statistical Quality Control*, 2001, 6, 207-227.

[68] T.Sukchotrat, S.B. Kim and F. Tsung, "One-class classification-based control charts for multivariate process monitoring", *IIE Transactions*, vol. 42, no. 2, pp. 107-120, 2009.

[69] R. Sun and F. Tsung, "A kernel distance-based multivariate control chart using support vector methods," *International Journal of Production Research*, vol. 41, no. 13, pp. 2975-2989, 2003.

[70] J.A.K. Suykens, J. Vandewalle, "Least Squares Support Vector Machine Classifiers", *Neural Processing Letters*, vol. 9, pp.293-300, 1999.

[71] Tax, D.M.J. (2001) One-class classification: concept-learning in the absence of counter-examples. PhD thesis, Delf University of Technology, Netherlands.

[72] D. M. J. Tax and R. P. W. Duin, "Uniform object generation for optimizing one-class classifiers," *Journal of Machine Learning Research*, vol. 2, pp. 251-256, 2002.

[73] D. M. J. Tax and R. P. W. Duin, "Data domain description using support vectors," *European Symposium on Aritficial Neural Networks*, pp. 251-256, 1999.

[74] W. H. Woodall and M. M. Ncube. Multivariate CUSUM quality control procedures. *Technometrics*, 1985.

[75] Yoo, C. K., Lee, J. M., Vanrolleghem, P. A. and Lee, I. B. (2004). On-line monitoring of batch processes using multiway independent component analysis. *Chemometrics and Intelligent Laboratory Systems*, 71(2), 151–163.

[76] Yuan, C. and Casasent, D., Support vector machines for class representation and discrimination, *in International Joint Conference on Neural Networks*, 2003, Portland, OR, 1610-1615.

[77] Yan-Zhong, Hong-Lie, Yan-Ju and Jin-Gang, (2014). Hybrid patterns recognition of control chart based on WA-PCA-PSO-SVM, International Journal of Control and Automation Vol.7, No.10, pp.91-98.

[78] C. Zou and F. Tsung, "A multivariate sign EWMA control chart," *Technometrics*, vol. 53, no. 1, pp. 84-97, 2011.