ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ

ΚΑΙ

ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

Εφαρμογή των SVM και k-NN τεχνικών σε προβλήματα Στατιστικού Ελέγχου Ποιότητας

Διπλωματική Εργασία της:

Αγγελικής Ηλία Ζαρωτιάδου

Επιβλέπων Καθηγητής:

Κουκουβίνος Χρήστος, Καθηγητής Ε.Μ.Π.

Αθήνα 2011

Περιεχόμενα:

Ευχαριστίες

Θα ήθελα να ευχαριστήσω από καρδιάς τον επιβλέποντα καθηγητή κ. Χρήστο Κουκουβίνο, για την δυνατότητα ανάληψης της επικείμενης διπλωματικής εργασίας.

Θα ήθελα να ευχαριστήσω την διδάκτορα Χριστίνα Πάρπουλα, για τις πληροφορίες και το υλικό που μου παρείχε, καθώς και για τις πολύτιμες συμβουλές και διορθώσεις της.

Ευγνώμων είμαι στους φίλους μου που μου συμπαραστάθηκαν επίσης σε αυτή την προσπάθεια αλλά κυρίως στην οικογένειά μου που μου έδειξε πως ακόμη και κάτω από αντίξοες συνθήκες υπάρχει δυνατότητα να φέρουμε σε πέρας τις υποχρεώσεις μας.

Εισαγωγή

Η χρήση των διαγραμμάτων στατιστικού ελέγχου ποιότητας είναι η ανίχνευση τυχόν αλλαγών στην μέση τιμή και τυπική απόκλιση των χαρακτηριστικών μιας διεργασίας. Όταν τα διαγράμματα ελέγχου χρησιμοποιούνται σωστά μετά από κατάλληλη επεξεργασία των δεδομένων, βοηθούν στην ανεύρεση των αιτιών που αλλοιώνουν την διεργασία, στην εξάλειψη αυτών των αιτιών και στην βελτίωση τελικά της ποιότητας. Σκοπός της παρούσας διπλωματικής εργασίας είναι η παρουσίαση των βασικών μεθόδων εξόρυξης δεδομένων και των παραμετρικών και μη παραμετρικών διαγραμμάτων ελέγχου.

Στην παρούσα εργασία περιλαμβάνονται επτά κεφάλαια τα οποία οργανώνονται ως εξής:

- Στο Κεφάλαιο 1 παρουσιάζεται η εξόρυξη δεδομένων, η συμβολή της στην στατιστική, η διαδικασία που ακολουθείται για να γίνει η εξόρυξη δεδομένων. Επίσης αναφέρονται οι βασικές έννοιες και ορισμοί που χρησιμοποιούνται για την διαδικασία. Το Data Mining χωρίζεται σε κατηγορίες οι οποίες παρουσιάζονται επαρκώς σε αυτό το κεφάλαιο. Τέλος, αναφέρονται οι τομείς στους οποίους έχει εφαρμογή το Data Mining.

- Στο Κεφάλαιο 2 παρουσιάζεται η μία από τις δύο βασικές μεθόδους εξόρυξης δεδομένων, η kNN. Περιγράφονται οι διάφοροι ταξινομητές μεταβλητής και σε ποια περίπτωση χρησιμοποιείται ο καθένας, καθώς και τα συμπεράσματα στα οποία οδήγησαν διάφορα πειράματα τα οποία περιγράφονται μαζί με τα αποτελέσματα τους.

- Στο Κεφάλαιο 3 παρουσιάζεται η δεύτερη μέθοδος εξόρυξης δεδομένων, η SVM. Μετά τον επεξήγηση της μεθόδου, ακολουθεί πλήρης εξήγηση του τρόπου αναπαράστασης, των ταξινομητών και των τεχνασμάτων τους πυρήνα. Τέλος, αναφέρεται η χρήση της μεθόδου στην παλινδρόμηση, τα πλεονεκτήμετα και μειονεκτήματα της μεθόδου καθώς και οι εφαρμογές της.

- Στο Κεφάλαιο 4 γίνεται η αναφορά στον στατιστικό έλεγχο ποιότητας, στην αποτελεσματικότητά του και η σύνδεση με τα διαγράμματα ελέγχου, καθώς και μια ιστορική αναδρομή του στατιστικού ελέγχου διεργασιών.

- Στο Κεφάλαιο 5 παρουσιάζονται τα βασικά διαγράμματα ελέγχου, τα οποία διακρίνονται σε διαγράμματα ελέγχου μεταβλητών, διαγράμματα ελέγχου ιδιοτήτων και αθροιστικά διαγράμματα ελέγχου. Τέλος γίνεται παρουσίαση των διαγραμμάτων ελέγχου με κινητούς μέσους και εκθετικά βάρη.

- Στο Κεφάλαιο 6 γίνεται παρουσίαση των παραμετρικών και μη παραμετρικών διαγραμμάτων ελέγχου, μετά από μια αναφορά στον στατιστικό έλεγχο

διεργασίας και το πώς συνδέεται με αυτά τα διαγράμματα. Παρουσιάζονται τα διαγράμματα που βασίζονται στον αλγόριθμο SVDD, στον αλγόριθμο kNNDD, ο τρόπος που γίνεται η μελέτη προσομοίωσης και η εφαρμογή των διαγραμμάτων.

- Στο Κεφάλαιο 7 παρουσιάζονται συνοπτικά τα συμπεράσματα που προέκυψαν από αυτή την εργασία και ο τρόπος που μπορεί να είναι χρήσιμη στην μελέτη των διαγραμμάτων ελέγχου κάτω από την κατάλληλη επεξεργασία των δεδομένων.

Abstract

The use of statistical quality control charts is to detect in the mean and standard deviation of the characteristics of a process. When control charts are used correctrly after proper processing of data, they help in the finding the causes affecting the process to eliminate these causes and improve final quality. The purpose of this thesis is tthe presentation of key data mining methods and parametric and non-parametric control charts.

The seven chapters are organised as follows:

- In Chapter 1 the data moning contribution to the stastistics is presented and the procedure for making the data mining. In addition, the basic concepts and definitions that are used in the process, are referred. Data mining is devided into categories which are presented adequately in this chapter. Finally, the areas which have have been implemented as data mining are indicated.

- In Chapter 2one of the main methods of data mining is presented, the kNN. There is a description of the different classifiers and what if each of them is used. Furthermore, the conclusions to which various experiments led are referred.

- In Chapter 3 another important method of data mining is presented, the SVM. The description of the method is followed by a full presentation of classifiers and tricks of the kernel. In the end, there is an axplanation about how the method is used in regression, as well as the advantages and disadvantages of this method.

- In Chapter 4 there is a reference to statistical quality control, the efficiency of it and the connectivity to the control charts. A historical overview of statistical process control is presented.

- In Chapter 5 the basic control charts are presented. They are divided into variable control charts, contol charts and cumulative control charts. There is also a presentation of control charts with moving averages and exponential weights.

- In Chapter 6 there is a presentation of the parametric and non-parametric control charts, after a reference to statistical process control and how it is related to these charts. Diagrams based on the SVDD algorithm are explaned, as well as the kNNDD algorithm. The simulation and the application of diagrams are also presented.

- In Chapter 7 there is a summary of the conclusions drawn from this work and how they can bu useful in the study of control charts under the appropriate data processing.

1 ΚΕΦΑΛΑΙΟ – ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

1.1 Εισαγωγικά στοιχεία για το Data Mining

Εξαιτίας της αύξησης του όγκου πληροφοριών και την μείωση της ποσότητας που γίνεται κατανοητή, δημιουργήθηκε η ανάγκη για νέες μεθόδους εξόρυξης πληροφοριών. Με αυτό το αντικείμενο ασχολείται το επιστημονικό πεδίο που καλούμε Data Mining, δηλαδή το σύνολο διαδικασίας εξόρυξης πληροφορίας και γνώσης από τον όγκο των δεδομένων. Η λέξη data, αναφέρεται ως συλλογή αριθμών ή συμβολοσειρών επεξεργάσιμων από υπολογιστές. Για την διαδικασία ανακάλυψης γνώσης σε βάσεις δεδομένων KDD (Knowledge Discovery in Databases) έχουν δείξει ενδιαφέρον οι επιστημονικές κοινότητες. Το Data Mining είναι ένας κλάδος συνεχώς εξελισσόμενος και σύμφωνα με το MIT Technology Review αποτελεί μια από τις 10 πιο ανερχόμενες τεχνολογίες που θα αλλάξουν τον κόσμο. Το 1995 έλαβε χώρα το πρώτο διεθνές συνέδριο για το Data Mining και KDD. Το data mining σε συνδυασμό με τις εφαρμοσμένες στατιστικές μεθόδους είναι απαραίτητα εργαλεία για την εξαγωγή γνώσεων από τέτοιες βάσεις δεδομένων.

Η μεθοδολογική έρευνα στον τομέα της στατιστικής έχει οδηγήσει στην ανάπτυξη διαδικασιών και μεθόδων που μπορούν να χρησιμοποιηθούν για την ανάλυση μεγάλων βάσεων δεδομένων. Πολλές από τις μεθοδολογίες που χρησιμοποιούμε στο data mining έχουν προέλθει από την έρευνα και την ανάπτυξη στο τμήμα της μηχανικής μάθησης (machine learning) και από το τμήμα της στατιστικής, κυρίως από την πολυμεταβλητή και υπολογιστική στατιστική. Η μηχανική μάθηση είναι συνδεδεμένη με τις επιστήμες των υπολογιστών και την τεχνητή νοημοσύνη και ασχολείται με την εύρεση σχέσεων (συσχετίσεων) και κανονικοτήτων (προτύπων) μεταξύ δεδομένων που να μπορούν να γενικευτούν. Στόχος της υπολογιστικής εκμάθησης είναι η αναπαραγωγή διαδικασιών που προέρχονται από την επεξεργασία των υπαρχόντων (παρατηρούμενων) δεδομένων και επιτρέπουν στους αναλυτές την γενίκευση αυτών και σε μη παρατηρούμενα δεδομένα. Ο όρος «ανακάλυψη γνώσης σε βάσεις δεδομένων» (Knowledge Discovery in Databases, KDD) επινοήθηκε για την περιγραφή όλων αυτών των μεθόδων που στόχο έχουν την εύρεση συσχετίσεων και συνθηκών μεταξύ των παρατηρούμενων δεδομένων. Τελικά ο όρος αυτός κατέληξε να περιγράφει την όλη διαδικασία υπολογισμού πληροφοριών από μια βάση δεδομένων, από τον ορισμό των αρχικών στόχων έως την εξαγόμενη γνώση που προκύπτει μετά την εφαρμογή των διάφορων τεχνικών του data mining.

Σύμφωνα με τον ορισμό του Fayaad το data mining, δηλαδή η ανακάλυψη γνώσης από βάσεις δεδομένων είναι μια ντετερμινιστική διαδικασία αναγνώρισης έγκυρων, καινοτόμων, πιθανόν χρήσιμων και κατανοητών προτύπων στα δεδομένα.
Ένας ευρέως αποδεκτός ορισμός του data mining είναι ο εξής:
Εξόρυξη δεδομένων καλείται η εξεύρεση (σημαντικών, αυτονόητων, άγνωστων και πιθανόν χρήσιμων) πληροφοριών ή επαναλαμβανόμενων Προτύπων (patterns) σε

τεράστιες βάσεις δεδομένων. Παραβλέποντας το γεγονός ότι είναι δύσκολο να καθοριστούν επακριβώς τα όρια μελέτης αυτού του επιστημονικού κλάδου μέσα σε έναν και μόνο ορισμό παραθέτουμε έναν γενικό, συνοπτικό ορισμό του data mining: Η **εξόρυξη δεδομένων (data mining)** είναι η διαδικασία που ακολουθείται από τη συλλογή, εξερεύνηση και μοντελοποίηση μεγάλων παρατηρούμενων συνόλων δεδομένων με σκοπό την εύρεση συνθηκών και συσχετισμών που θα οδηγήσουν σε εμφανή, κατανοητά και χρήσιμα αποτελέσματα για τους κατόχους των δεδομένων.

Σε ένα πρόβλημα data mining υπάρχει το σύνολο δεδομένων εκπαίδευσης (training set), του οποίου είναι γνωστή η τιμή του αποτελέσματος και των χαρακτηριστικών που μας ενδιαφέρουν και προσπαθούμε να κατασκευάσουμε ένα μοντέλο πρόβλεψης με βάση αυτά τα δεδομένα. Με την χρήση αυτού του μοντέλου θα προβλέψουμε τα αποτελέσματα νέων συνόλων δεδομένων εξέτασης (test set) των οποίων είναι γνωστές οι τιμές των χαρακτηριστικών αλλά όχι η τιμή του αποτελέσματος, δηλαδή η τιμή της τάξης. Το data mining στηρίζεται στην κατασκευή υπολογιστικών προγραμμάτων που χρησιμοποιούν στατιστικά αποτελέσματα για το κρησάρισμα των βάσεων δεδομένων και την εξαγωγή προτύπων και άλλων πληροφοριών. Οι σχέσεις που προκύπτουν από την διαδικασία του data mining αναφέρονται ως πρότυπα (patterns) τα οποία περιλαμβάνουν γραμμικές εξισώσεις, κανόνες, συστάδες (clusters), γραφήματα, δέντρα καθώς και επαναλαμβανόμενα πρότυπα σε χρονοσειρές. Τα πιθανά προβλήματα τα οποία μπορεί να προκύψουν στην διαδικασία επιλογής προτύπων είναι η δημιουργία προτύπων που δεν έχουν ενδιαφέρον, είναι ξεπερασμένα ή πολύπλοκα. Επίσης είναι πιθανό το πρότυπο να είναι αποτέλεσμα συμπτώσεων της συγκεκριμένης βάσης δεδομένων, ή τα πραγματικά δεδομένα να είναι ελλιπή ή διαστρεβλωμένα και να προκύπτουν ανακριβή συμπεράσματα. Για αυτό το λόγο οι αλγόριθμοι εξόρυξης δεδομένων πρέπει να είναι ανθεκτικοί ώστε να ανταποκρίνονται σε μη τέλεια δεδομένα και να εξάγουν κανόνες χρήσιμους αν όχι ακριβείς. Η εύρεση ισχυρών προτύπων είναι ένα πολύ χρήσιμο εργαλείο για την ακριβή πρόβλεψη μελλοντικών δεδομένων, για την γενίκευση από ένα δείγμα του συνόλου στο πλήρες σύνολο αλλά και για την συμπίεση μεγάλων δεδομένων σε μικρότερα ώστε να γίνουν κατανοητά και χρήσιμα. Η εξόρυξη δεδομένων είναι ένα διεπιστημονικό πεδίο που φέρει κοντά τεχνικές από τη μηχανική μάθηση, την αναγνώριση προτύπων, τη στατιστική και τις βάσεις δεδομένων, με σκοπό την οπτικοποίηση, έτσι ώστε να αντιμετωπιστεί το ζήτημα της εξαγωγής πληροφορίας από μεγάλες βάσεις δεδομένων.

1.2 Το Data mining στη συμβολή των πεδίων της Στατιστικής και της Μηχανικής Μάθησης

Ο συνδυασμός του data mining με μεθόδους στατιστικής, εκμάθησης μηχανής, αλγορίθμους ομαδοποίησης και μεθόδους οπτικοποίησης συμβάλλει στην εξαγωγή ασφαλών αποτελεσμάτων σε πρακτικά προβλήματα. Τα βασικά συστατικά του data mining για την ανάλυση δεδομένων είναι η στατιστική ανάλυση και οι αλγόριθμοι

εξόρυξης πληροφορίας. Το πλεονεκτήματα της στατιστικής είναι η μαθηματική ερμηνεία των δεδομένων και βασικό μειονέκτημα της αποτελεί η χρήση μικρών συνόλων δεδομένων. Η επιστήμη της Στατιστικής ήταν πάντα σχετική με τη δημιουργία μεθόδων για την ανάλυση δεδομένων. Η κύρια διαφορά μεταξύ των στατιστικών μεθόδων και των μεθόδων μηχανικής μάθησης είναι ότι οι στατιστικές μέθοδοι αναπτύσσονται συνήθως σε σχέση με τα δεδομένα που αναλύθηκαν και σύμφωνα με ένα εννοιολογικό πρότυπο αναφοράς.

Για μεγάλο χρονικό διάστημα οι στατιστικολόγοι θεωρούσαν την εξόρυξη δεδομένων ως συνώνυμο με το «αλιεία δεδομένων» ή «δεδομένα snooping». Σε όλες αυτές τις περιπτώσεις η εξόρυξη δεδομένων είχε αρνητικές συνδηλώσεις. Η ιδέα αυτή προέκυψε λόγω των δύο βασικών επικρίσεων. Πρώτον, δεν υπάρχει μόνο ένα θεωρητικό μοντέλο αναφοράς, αλλά πολλά μοντέλα που βρίσκονται σε ανταγωνισμό μεταξύ τους. Αυτά τα μοντέλα έχουν επιλεγεί ανάλογα με τα δεδομένα που εξετάζονται κάθε φορά. Η κριτική αυτής της διαδικασίας είναι ότι είναι πάντοτε δυνατό να βρούμε ένα μοντέλο, ανεξαρτήτως πολυπλοκότητας, το οποίο θα προσαρμοστεί καλά στα δεδομένα. Δεύτερον, το μεγάλο ποσό των διαθέσιμων δεδομένων μπορεί να οδηγήσει σε ανύπαρκτες σχέσεις που βρέθηκαν μεταξύ των δεδομένων.

Οι σύγχρονες μέθοδοι εξόρυξης δεδομένων αποδίδουν ιδιαίτερη σημασία στη δυνατότητα γενίκευσης των αποτελεσμάτων. Αυτό σημαίνει ότι όταν επιλέγουμε ένα μοντέλο, η επίδοση πρόβλεψης λαμβάνεται υπόψη και τα πιο πολύπλοκα μοντέλα δεν παραλείπονται. Είναι δύσκολο να αγνοήσουμε το γεγονός ότι πολλά σημαντικά ευρήματα δεν είναι γνωστά εκ των προτέρων και δεν μπορούν να χρησιμοποιηθούν στην ανάπτυξη μιας υπόθεσης της έρευνας. Αυτό συμβαίνει κυρίως όταν υπάρχουν μεγάλες βάσεις δεδομένων.

Οι μεγάλες βάσεις δεδομένων είναι ένα από τα χαρακτηριστικά που διακρίνει την εξόρυξη δεδομένων από τη στατιστική ανάλυση. Λαμβάνοντας υπόψη ότι η στατιστική ανάλυση ασχολείται με την ανάλυση των πρωτογενών δεδομένων που έχουν συγκεντρωθεί για να ελέγξει τις υποθέσεις μιας συγκεκριμένης έρευνας, η εξόρυξη δεδομένων μπορεί να ασχολείται επίσης με δευτερογενή δεδομένα που συλλέγονται για άλλους λόγους. Επιπλέον, τα στατιστικά στοιχεία μπορεί να είναι πειραματικά δεδομένα (ίσως το αποτέλεσμα ενός πειράματος που κατανέμει τυχαία όλα τα στατιστικά δεδομένα για διάφορα είδη θεραπείας), αλλά στην εξόρυξη δεδομένων τα δεδομένα είναι συνήθως παρατηρούμενα δεδομένα.

Οι Berry και Linoff (1997) διακρίνουν δύο αναλυτικές προσεγγίσεις για την εξόρυξη δεδομένων. Διαφοροποιούν την από την κορυφή προς τα κάτω (top-down) ανάλυση (επιβεβαίωσης) και την κάτω προς τα πάνω (bottom-up) ανάλυση (διερευνητική). Η top-down ανάλυση έχει ως στόχο να επιβεβαιώσει ή να απορρίψει υποθέσεις και προσπαθεί να διευρύνει τις γνώσεις μας για ένα εν μέρει κατανοητό φαινόμενο. Αυτό επιτυγχάνεται κυρίως με τη χρήση των παραδοσιακών στατιστικών μεθόδων. Bottom-

up ανάλυση έχουμε όταν ο χρήστης αναζητά χρήσιμες πληροφορίες, που δεν είχαν παρατηρηθεί στο παρελθόν, ερευνώντας τα δεδομένα και αναζητώντας τρόπους για τη σύνδεσή τους με τη δημιουργία υποθέσεων. Η bottom-up προσέγγιση είναι χαρακτηριστικό της εξόρυξης δεδομένων. Στην πραγματικότητα, οι δύο προσεγγίσεις είναι συμπληρωματικές. Οι πληροφορίες που λαμβάνονται από την bottom-up ανάλυση, η οποία προσδιορίζει σημαντικές σχέσεις και τάσεις, δεν μπορούν να εξηγήσουν γιατί αυτές οι ανακαλύψεις είναι χρήσιμες και σε ποιο βαθμό είναι έγκυρες. Τα εργαλεία επιβεβαίωσης της top-down ανάλυσης μπορούν να χρησιμοποιηθούν για να επιβεβαιώσουν τις ανακαλύψεις και να αξιολογήσουν την ποιότητα των αποφάσεων που βασίζονται σε αυτές τις ανακαλύψεις.

Υπάρχουν τουλάχιστον τρία άλλα σημεία που διαφοροποιούν τη στατιστική ανάλυση δεδομένων από την εξόρυξη δεδομένων. Πρώτον, η εξόρυξη δεδομένων αναλύει μεγάλες μάζες δεδομένων και αυτό συνεπάγεται νέες εκτιμήσεις για την στατιστική ανάλυση. Για πολλές εφαρμογές, είναι αδύνατη η ανάλυση ή ακόμη και η πρόσβαση σε ολόκληρη τη βάση δεδομένων για λόγους μη αποδοτικότητας του υπολογιστή. Ως εκ τούτου, καθίσταται απαραίτητο να λαμβάνουμε και να εξετάζουμε ένα δείγμα από τη βάση δεδομένων. Σε αυτή τη δειγματοληψία πρέπει να λαμβάνονται υπόψη οι στόχοι της εξόρυξης δεδομένων, και αυτό δεν μπορεί να συμβεί χρησιμοποιώντας τις παραδοσιακές μεθόδους στατιστικής θεωρίας. Δεύτερον, πολλές βάσεις δεδομένων δεν οδηγούν σε κλασικές μορφές οργάνωσης των στατιστικών δεδομένων, όπως για παράδειγμα, τα δεδομένα που προέρχονται από το διαδίκτυο. Αυτό δημιουργεί την ανάγκη για κατάλληλες αναλυτικές μεθόδους έξω από το πεδίο των στατιστικών. Τρίτον, τα αποτελέσματα της εξόρυξης δεδομένων πρέπει να έχουν κάποια συνέπεια. Αυτό σημαίνει ότι πρέπει να δοθεί προσοχή στα αποτελέσματα που επιτεύχθηκαν στις επιχειρήσεις με τα μοντέλα ανάλυσης δεδομένων.
Σε μία σύγκριση μεταξύ data mining και στατιστικής παρουσιάζουμε συνοπτικά τα παρακάτω:

- Data Mining
    - Δεν χρειάζονται υποθέσεις.
    - Ικανότητα εύρεσης προτύπων σε μεγάλες ποσότητες δεδομένων.
    - Χρησιμοποίηση όλων των διαθέσιμων δεδομένων.
    - Ορολογία που χρησιμοποιείται: πεδίο, καταχώρηση, μάθηση με επίβλεψη, μάθηση χωρίς επίβλεψη.

- Στατιστική
    - Μη κατάλληλες τεχνικές για μεγάλα σύνολα δεδομένων.
    - Βασισμένη στην δειγματοληψία.
    - Χρήση τεστ υποθέσεων (hypothesis testing)
    - Ορολογία που χρησιμοποιείται: μεταβλητή, παρατήρηση, ανάλυση της εξάρτησης, ανάλυση της αλληλεξάρτησης.

Ο συνδυασμός εφαρμογής αλγορίθμων σε υποδείγματα (μηχανική μάθηση) και στατιστικής ανάλυσης παρέχει αξιόπιστα και ακριβή αποτελέσματα στις τεχνικές ανάλυσης δεδομένων. Ο όρος αντίληψη (concept) αναφέρεται στο αντικείμενο της μάθησης και ο στόχος του data mining είναι η εύρεση κατανοητής και λειτουργικής περιγραφής μιας αντίληψης με εφαρμογή κατάλληλων και αποτελεσματικών αλγορίθμων.

Τα απαιτούμενα βήματα που πρέπει να ακολουθήσει ένας αναλυτής για την ανάπτυξη αποτελεσματικών αλγορίθμων συνοψίζονται στα εξής:

1. Κατασκευή ξεχωριστών αλγορίθμων για κάθε περίπτωση, εξαιτίας της ποικιλίας των τύπων δεδομένων και των στόχων του data mining.
2. Κατασκευή αποτελεσματικών αλγορίθμων με δυνατότητα κλιμάκωσης σε μεγάλη βάση δεδομένων σε αποδεκτό και αναμενόμενο χρόνο.
3. Κατασκευή αλγορίθμων ικανών να διαχειρίζονται τον θόρυβο και τα δεδομένα εξαιρέσεις.
4. Κατασκευή αλγορίθμων όπου η μη τελειότητα των αποτελεσμάτων εκφράζεται μέσα από μέτρα αβεβαιότητας.

Αυτά τα βήματα για την κατασκευή στατιστικών μοντέλων και εργαλείων, οδηγούν τον πειραματιστή στην μέτρηση της ποιότητας της ανακαλυφθείσας γνώσης, του ενδιαφέροντος που παρουσιάζει και της αξιοπιστίας της.

1.3 Διαδικασία εξόρυξης δεδομένων

Παρουσιάζουμε αναλυτικά τα εφτά κύρια στάδια της διαδικασίας εξόρυξης δεδομένων:

i. Ορισμός των στόχων και των προς ανάλυση αντικειμένων:

Καθορισμός των στόχων της ανάλυσης.
Σαφής διατύπωση του προβλήματος.
Σαφείς στόχοι που πρέπει να επιτευχθούν.
Οργάνωση των μεθόδων που θα χρησιμοποιήσουμε.
Ανάπτυξη και κατανόηση του στόχου της περιοχής της εφαρμογής, της προγενέστερης γνώσης του τομέα που εξετάζεται και τους στόχους του τελικού χρήστη.

ii. Επιλογή, οργάνωση και επισκόπηση των δεδομένων:

Επιλογή δεδομένων που θα χρησιμοποιήσουμε για την ανάλυση.
Ορισμός των πηγών των δεδομένων-συνήθως πηγή είναι η εταιρία που κάνει την ανάλυση για λόγους οικονομίας και επειδή είναι αποτελέσματα πειραμάτων των διαδικασιών της ίδιας της εταιρίας.
Δημιουργία data marts: οδηγεί σε παρουσίαση των δεδομένων σε μορφή πινάκων που βασίζεται στις ανάγκες της ανάλυσης και τους ήδη καθορισμένους στόχους. Έχοντας τον πίνακα δεδομένων είναι εύκολο να κάνουμε ποιοτικό έλεγχο στα διαθέσιμα

δεδομένα και να αποφασίσουμε αν κάποια μεταβλητή είναι ακατάλληλη ή αν κάποια πληροφορία απουσιάζει.
Μικρό δείγμα δεδομένων: μειώνει το χρόνο ανάλυσης και το υπολογιστικό κόστος.
Έλεγχος εγκυρότητας του μοντέλου
Μείωση του ρίσκου προσομοίωσης της στατιστικής μεθόδου σε μη κανονικοποίηση και δεν υπάρχει κίνδυνος να χαθεί η ικανότητα της γενίκευσης και πρόγνωσης του μοντέλου.
Τα διαφορετικά είδη αποθηκών πληροφοριών που χρησιμοποιούνται στην διαδικασία εξόρυξης δεδομένων παρέχουν την δυνατότητα συνδυασμού των πηγών δεδομένων για τον καθορισμό του συνόλου που θα χρησιμοποιηθεί τελικά στην διαδικασία εξόρυξης. Με αυτόν τον τρόπο καθορίζεται ο στόχος – σύνολο δεδομένων και επιλέγεται το σύνολο δεδομένων στο οποίο θα εφαρμοστεί η διαδικασία εξόρυξης.

iii. Διερευνητική ανάλυση δεδομένων:

Προκαταρκτική ανάλυση δεδομένων: μετατροπή των αρχικών μεταβλητών για την καλύτερη κατανόηση ή στατιστικές μέθοδοι που βασίζονται στην ικανοποίηση των αρχικών υποθέσεων.
Εντοπισμός στοιχείου που διαφέρει σημαντικά από τα υπόλοιπα.
Πρόβλεψη της κατάλληλης στατιστικής μεθόδου για την επόμενη φάση της ανάλυσης.
Η διερευνητική ανάλυση μπορεί αν υποδηλώνει την ανάγκη για νέα εξόρυξη δεδομένων αν τα υπάρχοντα θεωρούνται ανεπαρκή.
Μείωση των δεδομένων, αν είναι απαραίτητο, και διαχωρισμός των δεδομένων σε: δεδομένα εκπαίδευσης (training set), δεδομένα επαλήθευσης-επικύρωσης (quiz set-validation) και δεδομένα ελέγχου-εξέτασης (test set). Μετασχηματισμός ή παγίωση των δεδομένων σε μορφές κατάλληλες για την διαδικασία εξόρυξης. Με την χρήση μεθόδων μείωσης διαστάσεων ή μετασχηματισμού γίνεται μείωση του αριθμού των υπό εξέταση μεταβλητών και εύρεση του κατάλληλου συνόλου δεδομένων.
Προσδιορισμός του είδους μάθησης του data-mining και εξόρυξη δεδομένων: ταξινόμηση-πρόβλεψη-ομαδοποίηση.

iv. Καθορισμός των στατιστικών μεθόδων:

Ταξινόμηση των υφιστάμενων μεθόδων.
Επιλογή της μεθόδου ανάλογα με το πρόβλημα και το είδος των διαθέσιμων δεδομένων. Επιλογή των τεχνικών του data-mining που θα χρησιμοποιηθούν.

Τρεις κύριες κατηγορίες μεθόδων:

- Περιγραφικές μέθοδοι

Περιγράφουν τις ομάδες των δεδομένων πιο σύντομα. Καλούνται επίσης συμμετρικές, χωρίς επίβλεψη, έμμεσες. Όλες οι διαθέσιμες μεταβλητές αντιμετωπίζονται στο ίδιο επίπεδο και δεν υπάρχουν σχέσεις αιτιότητας.

- Μέθοδοι πρόβλεψης

Περιγράφουν μία ή περισσότερες μεταβλητές σε σχέση με τις υπόλοιπες. Καλούνται επίσης ασύμμετρες, υπό επίβλεψη, άμεσες. Εξετάζουν την ύπαρξη

κανόνων ταξινόμησης ή πρόβλεψης βάσει των δεδομένων. Οι κυριότερες μέθοδοι αυτού του είδους είναι του τομέα της μηχανικής μάθησης, όπως τα νευρωνικά δίκτυα, τα δέντρα αποφάσεων, καθώς και τα γραμμικά και λογιστικά μοντέλα παλινδρόμησης.

- Τοπικές μέθοδοι

Εντοπίζουν ιδιαίτερα χαρακτηριστικά που σχετίζονται με ένα υποσύνολο της βάσης δεδομένων. Οι περιγραφικές και οι μέθοδοι πρόβλεψης δεν είναι τοπικές μέθοδοι, είναι γενικές. Τοπικές μέθοδοι είναι: οι κανόνες συσχέτισης για την ανάλυση των συναλλαγών (transactional data) των δεδομένων και την ταυτοποίηση των ανώμαλων παρατηρήσεων (outliers). Η κατάταξη αυτή είναι εξαντλητική από λειτουργική άποψη και κάθε μέθοδος μπορεί να χρησιμοποιηθεί μόνη της ή ως στάδιο σε ανάλυση με πολλαπλά στάδια.

v. Ανάλυση των στοιχείων:

Μετάφραση των στατιστικών μεθόδων σε κατάλληλους αλγορίθμους για την σύνθεση των αποτελεσμάτων από την διαθέσιμη βάση δεδομένων. Για τις τυπικές μεθόδους δεν είναι απαραίτητη η ανάπτυξη ειδικών αλγορίθμων, αλλά θα πρέπει να είναι επαρκείς. Ερμηνεία των αποτελεσμάτων κατά την λήψη αποφάσεων.

vi. Αξιολόγηση των στατιστικών μεθόδων:

Πρέπει να επιλέξουμε το καλύτερο μοντέλο ανάλυσης δεδομένων. Αυτό θα γίνει βάσει των αποτελεσμάτων από τις μεθόδους. Με την χρήση κάποιων μέτρων γίνεται η αξιολόγηση των προτύπων και των μοτίβων ώστε να προσδιοριστούν εκείνα που αντιπροσωπεύουν καλύτερα την γνώση και είναι εκείνα που μας ενδιαφέρουν.

Αυτό το στάδιο αποτελεί ένα σημαντικό διαγνωστικό έλεγχο για την εγκυρότητα των στατιστικών μεθόδων. Αν οι μέθοδοι που επιλέξαμε δεν επιτρέπουν την επίτευξη των στόχων μας πρέπει να ορίσουμε μια νέα μέθοδο για την ανάλυση και να ξεκινήσουμε την διαδικασία από την αρχή.

Στην αξιολόγηση της μεθόδου πρέπει να λαμβάνουμε υπόψη τον περιορισμένο χρόνο, τους περιορισμένους πόρους, την ποιότητα των δεδομένων, και τα διαθέσιμα δεδομένα. Στο data mining χρησιμοποιούνται συνήθως παραπάνω από μία στατιστικές μέθοδοι, για να είναι ολοκληρωμένη η εικόνα του προβλήματος.

Για την επιλογή του καλύτερου μοντέλου συγκρίνουμε τα αποτελέσματα που παίρνουμε από τις διαφορετικές τεχνικές και αξιολογούμε τους κανόνες που δημιουργούνται.

vii. Ανάπτυξη του μοντέλου-Σταθεροποίηση της γνώσης-Παρουσίαση της γνώσης:

Η γνώση που έχει αποκτηθεί από την εξόρυξη ενσωματώνεται στο σύστημα και με την χρήση κάποιων τεχνικών αντιπροσώπευσης παρουσιάζεται στον χρήστη.

viii. Εφαρμογή των μεθόδων:

Το data mining παίζει σημαντικό ρόλο στην λήψη αποφάσεων μιας εταιρίας. Μετά την επιλογή και τον έλεγχο του μοντέλου σε ένα σύνολο δεδομένων, οι κανόνες

εφαρμόζονται στο σύνολο των διαθέσιμων δεδομένων. Η διαδικασία ενσωμάτωσης του data mining στην οργάνωση της εταιρίας αποτελείται από τις ακόλουθες τέσσερις φάσεις:

- Στρατηγική φάση: Μελέτη των διαδικασιών των επιχειρήσεων ώστε να προσδιοριστεί ο τομέας που θα ωφεληθεί περισσότερο από το data mining.
- Φάση κατάρτισης: Αξιολόγηση της λειτουργίας του data mining. Δημιουργία πιλοτικού έργου (project) και αξιολόγηση των αποτελεσμάτων με βάση τους στόχους και τα κριτήρια της στρατηγικής φάσης. Το πιλοτικό έργο πρέπει να είναι εύκολο στην χρήση. Αν είναι θετικό, τα πιθανά αποτελέσματα είναι δύο:
  Α. προκαταρκτική αξιολόγηση της χρησιμότητας των διάφορων τεχνικών του data mining.
  Β. ορισμός ενός πρωτότυπου συστήματος data mining.
- Δημιουργική φάση:
  Αν η αξιολόγηση των αποτελεσμάτων του πιλοτικού έργου για την εφαρμογή ενός πλήρους συστήματος data mining είναι θετική, τότε θα πρέπει να αναδιοργανωθούν τα δεδομένα, να αναπτυχθεί το πρωτότυπο μοντέλο μέχρι την ύπαρξη της πρώτης επιχειρησιακής έκδοσης και να επενδυθεί χρόνος για την υλοποίηση του έργου.
- Μεταναστευτική φάση
  Προετοιμασία ώστε να ενσωματωθεί επιτυχώς η διαδικασία εξόρυξης δεδομένων. Αυτό θα επιτευχθεί με την διδασκαλία πιθανών χρηστών και με συνεχή αξιολόγηση των αποτελεσμάτων των δεδομένων εξόρυξης.

Τα παραπάνω βήματα της διαδικασίας εξόρυξης δεδομένων αντιστοιχούν συνοπτικά στα βήματα SEMMA, μια μεθοδολογία data mining που αναπτύσσεται από το SAS:
SAMPLE : Δείγμα από το σύνολο δεδομένων. Διαχωρισμός των δεδομένων σε δεδομένα εκπαίδευσης (training set), δεδομένα επαλήθευσης-επικύρωσης (quiz set-validation) και δεδομένα ελέγχου-εξέτασης (test dataset).
EXPLORE : Εξέταση του συνόλου δεδομένων στατιστικά και γραφικά.
MODIFY : Μετασχηματισμός των μεταβλητών και απόδοση των ελλειπουσών τιμών.
MODEL : Εφαρμογή των προγνωστικών μοντέλων.
ACCESS: Σύγκριση των μοντέλων με χρήση ενός συνόλου δεδομένων επαλήθευσης.
Το SPSS-Clementine έχει παρόμοια μεθοδολογία, το μοντέλο CRISP-DM (Cross-Industry Standard Process for Data Mining).

1.4 Έννοιες και ορισμοί του data mining

i. Διαχείριση Δεδομένων

Η διαχείριση δεδομένων είναι μια διαδικασία ανάπτυξης αρχιτεκτονικών στοιχείων, πρακτικών και άλλων εφαρμογών σχετικά με τα δεδομένα. Η διαχείριση των δεδομένων περιλαμβάνει τα στάδια της μοντελοποίησης και της αποθήκευσης των δεδομένων, τα δεδομένα κίνησης, την διαχείριση της βάσης δεδομένων και την εξόρυξη των δεδομένων. Με την μοντελοποίηση των δεδομένων επιχειρείται η

δημιουργία δομής για τα δεδομένα έτσι ώστε να είναι προσιτά και αποδοτικά για την αποθήκευσή τους και την εξόρυξή τους για τις εκθέσεις και τις αναλύσεις. Για την δημιουργία δομής δεδομένων χρειάζεται μια σχέση με άλλα δεδομένα που να είναι σε τάξη. Εντός κάθε κατηγορίας τα δεδομένα μπορούν να ταξινομηθούν ανάλογα.

ii. Τύποι Δεδομένων

Τα δεδομένα τα αντλούμε από βάσεις δεδομένων του οργανισμού ή της εταιρίας που ενδιαφέρεται για την εφαρμογή του data mining. Τα δεδομένα αυτά είναι οργανωμένα υπό μορφή πινάκων. Οι πίνακες δεδομένων είναι σύνολα δεδομένων για ένα συγκεκριμένο σύνολο αντικειμένων και χαρακτηριστικών που έχουν οργανωθεί σε έναν *nxp* πίνακα. Οι γραμμές του πίνακα είναι οι τιμές του χαρακτηριστικού-μεταβλητής για ένα αντικείμενο και οι στήλες είναι οι τιμές ενός χαρακτηριστικού για κάθε αντικείμενο, δηλαδή το σύνολο των μετρήσεων *p* για κάθε αντικείμενο *n*. Τα χαρακτηριστικά-μεταβλητές είναι μια τιμή των αντικειμένων σε μια κλάση. Τα χαρακτηριστικά διακρίνονται σε ονομαστικά (nominal), τακτικά (ordinal), αριθμητικά (numeric), περιοδικά (intervals), συνεχή (continuous), διακριτά (discrete) και αναλογικά (ratio).

Οι τιμές των ονομαστικών (nominal) ή κατηγορικών χαρακτηριστικών είναι οι τίτλοι των αντικειμένων (ονόματα κλάσης ή κατηγοριών). Για αυτά τα χαρακτηριστικά δεν υπάρχει σχέση διάταξης ή απόστασης μεταξύ των τιμών τους. Τα δυαδικά χαρακτηριστικά αποτελούν μια ειδική περίπτωση των ονομαστικών χαρακτηριστικών, που παίρνουν δύο τιμές (ναι/όχι, σωστό/λάθος).
Οι τιμές των τακτικών χαρακτηριστικών (ordinal) μπορούν να διαταχθούν και η διάταξή τους έχει νόημα στην ανάλυση.
Οι τιμές των αριθμητικών χαρακτηριστικών (numeric) είναι αριθμοί που ανήκουν στο σύνολο ℤ και μπορεί να είναι συνεχή ή διακριτά.
Ο λόγος της απόστασης μεταξύ δύο τιμών περιοδικών χαρακτηριστικών (intervals) έχει νόημα για την ανάλυση. Το μηδέν και η μονάδα είναι τυπικά. Παραδείγματα τέτοιων χαρακτηριστικών είναι η θερμοκρασία και ο χρόνος.
Στα αναλογικά χαρακτηριστικά η μονάδα είναι συμβατική. Οι πράξεις μεταξύ των τιμών και η διάταξη των τιμών έχουν νόημα για την ανάλυση. Παραδείγματα τέτοιων χαρακτηριστικών είναι η μάζα και οι τιμές των προϊόντων.
Το χαρακτηριστικό μεταβλητή ως προς το οποίο εξετάζουμε την επίδραση των υπολοίπων χαρακτηριστικών καλείται τάξη ή κλάση και αποτελεί την απόκριση της ανάλυσής μας.

1.5 Κατηγοριοποίηση του data mining-Είδη μάθησης

Το data mining διακρίνεται σε δύο κατηγορίες:

A. Το κατευθυνόμενο (directed) data mining, που εξηγεί ή ταξινομεί κάποια πεδία στόχους.

B. Το μη κατευθυνόμενο(undirected) data mining, που βρίσκει ομοιότητες σε ομάδες εγγραφών χωρίς την χρήση συγκεκριμένου πεδίου στόχου ή συλλογής προκαθορισμένων κλάσεων.

Η γνώση που προκύπτει από μια διαδικασία εξόρυξης πληροφοριών από δεδομένα κατηγοριοποιείται ανάλογα με τον στόχο του προβλήματος που εξετάζουμε. Τα κύρια είδη μάθησης που διακρίνουμε είναι:

i. Ταξινόμηση (Classification) : η ταξινόμηση των υποδειγμάτων σε μια προκαθορισμένη τάξη (class) , δηλαδή η πρόβλεψη διακριτής κατηγορίας.
ii. Συσχέτιση (Association) : ο εντοπισμός συσχετίσεων μεταξύ των χαρακτηριστικών του συνόλου δεδομένων.
iii. Συσταδοποίηση (Clustering) / Ομαδοποίηση των δεδομένων : η εύρεση ομάδων όμοιων αντικειμένων και εκχώρηση υποδειγμάτων στις ομάδες αυτές, δηλαδή η ανάδειξη ομάδων όμοιων υποδειγμάτων.
iv. Αριθμητική πρόβλεψη : η πρόβλεψη μιας αριθμητικής ποσότητας. Η αριθμητική πρόβλεψη είναι όμοια με την ταξινόμηση με την διαφορά ότι εδώ η τάξη είναι αριθμητική.

Επίσης, τα είδη μάθησης διακρίνονται σε είδη με επίβλεψη (supervised learning) και χωρίς επίβλεψη (unsupervised learning). Οι μέθοδοι λοιπόν του data mining διακρίνονται σε δύο κατηγορίες:

A. Αλγόριθμοι εκμάθησης με επίβλεψη (supervised learning algoritghms), που χρησιμοποιούνται στην ταξινόμηση και την πρόβλεψη. Αυτή η μέθοδος χρειάζεται διαθέσιμα δεδομένα με γνωστή την τιμή του αποτελέσματος. Ο όρος supervised learning αναφέρεται στην διαδικασία του να τροφοδοτήσεις έναν αλγόριθμο με εγγραφές στις οποίες μια μεταβλητή απόκρισης (output variable) είναι γνωστή και ο αλγόριθμος να μάθει πώς να προβλέψει την τιμή με νέες εγγραφές όπου το αποτέλεσμα είναι άγνωστο. Δηλαδή γίνεται μοντελοποίηση της μεταβλητής απόκρισης (output variable) με βάση μία ή περισσότερες επεξηγηματικές μεταβλητές (input variable). Μερικά παραδείγματα τεχνικών με επίβλεψη είναι οι: neural networks, rule induction (decision trees), linear regression, logistic regression.
B. Αλγόριθμοι εκμάθησης χωρίς επίβλεψη (unsupervised learning algorithms), που χρησιμοποιούνται όταν δεν υπάρχει μία μεταβλητή απόκρισης να προβλεφθεί ή να ταξινομηθεί. Ο όρος usupervised learning αναφέρεται στην ανάλυση που γίνεται για την πληροφορία πέρα από την πρόβλεψη της τιμής μιας μεταβλητής που μας ενδιαφέρει. Δηλαδή αυτές οι τεχνικές χρησιμοποιούνται όταν δεν υπάρχει κάποιο πεδίο να προβλεφθεί αλλά εξερευνούνται οι σχέσεις μεταξύ των των δεδομένων ώστε να είναι γνωστή η δομή τους. Μερικά παραδείγματα τέτοιων τεχνικών είναι οι: kihonen network, two-step, k-mean.

Ανάλογα με το πρόβλημα, το σκοπό του data mining, την φύση των διαθέσιμων δεδομένων και τις δυνατότητες και τις προτιμήσεις του data miner επιλέγεται ένας κατάλληλος συνδυασμός supervised και usupervised τεχνικών.

1.6 Εφαρμογές του data mining

Ο τομέας που εξυπηρετείται περισσότερο από την εξόρυξη δεδομένων είναι ο χρηματοπιστωτικός καθώς και οι τηλεπικοινωνίες όπου ανιχνεύεται η δόλια χρήση των υπηρεσιών που προσφέρουν. Το data mining έχει συμβάλλει σημαντικά στον έλεγχο του κόστους και στην αύξηση των εσόδων. Η Ιατρική, η φαρμακοποιία και η Βιοϊατρική είναι επίσης τομείς που εξυπηρετούνται από την εφαρμογή του data mining. Η εξόρυξη δεδομένων χρησιμοποιείται από τις εταιρίες που ασχολούνται με τα οικονομικά για τον ορισμό των χαρακτηριστικών της αγοράς και την πρόβλεψη της απόδοσης μετοχής. Στον τομέα της ιατρικής η εφαρμογή του data mining κάνει εφικτή την παροχή επιστημονικών αποφάσεων για την διάγνωση και τη θεραπεία μιας ασθένειας με την σωστή διαχείριση των πληροφοριών από την βάση δεδομένων του νοσοκομείου για την πρόβλεψη της αποτελεσματικότητας των χειρουργικών επεμβάσεων, των εξετάσεων, των φαρμακευτικών αγωγών, αναπτύσσοντας έτσι την τηλεϊατρική.
Μερικά χαρακτηριστικά πεδία που εφαρμόζεται το data mining συνοψίζονται στα παρακάτω:

1. Ανάλυση εταιριών και διαχείριση ρίσκου:
    - Προβλέψεις
    - Διατήρηση πελατολογίου
    - Βελτιωμένη χρηματοδότηση

Π.χ. Δέντρα αποφάσεων από ιστορικά στοιχεία τραπεζών ώστε να προσδιορίζεται η δυνατότητα να δοθεί δάνειο σε έναν υποψήφιο πελάτη.

2. Ανάλυση αγοράς και διαχείριση:
    - Target marketing
    - Customer relation Management
    - Market basket analysis (supermarket)
    - Cross selling

Π.χ. Στις τράπεζες : έλεγχος ποιότητας και ανάλυση ανταγωνιστικότητας
Η περίπτωση «Diapers and beer» : Η παρατήρηση ότι όσοι αγοράζουν πάνες αγοράζουν και μπύρα οδηγεί στην τοποθέτησή τους σε κοντινά ράφια. Αν τοποθετηθεί και ένα άλλο προϊόν ανάμεσα, αυξάνονται τελικά οι πωλήσεις και στα 3 είδη.

3. Εντοπισμός απάτης και διαχείριση ρίσκου
    - Εξόρυξη κειμένου και web analysis
    - Ευφυείς απαντήσεις σε ερωτήματα

Π.χ. Άτομα που σκηνοθετούν ατυχήματα για να εισπράξουν τα χρήματα από την ασφαλιστική, άτομα που κλέβουν παρόχους τηλεπικοινωνιών, ή ο εντοπισμός ακατάλληλων ιατρικών μεθόδων.

2 ΚΕΦΑΛΑΙΟ – Η ΜΕΘΟΔΟΣ kNN

2.1 Εισαγωγή για την μέθοδο του k- Nearest Neighbor (kNN)

Η μέθοδος του k-πλησιέστερου γείτονα (k- Nearest Neighbor) είναι μια ευρέως χρησιμοποιούμενη τεχνική που έχει εφαρμογή στην ομαδοποίηση και την ταξινόμηση. Σε προβλήματα ταξινόμησης έχουν προταθεί μετασχηματισμοί της μεθόδου του πλησιέστερου γείτονα οι οποίοι εκμεταλλεύονται πληροφορίες από την δομή του συνόλου δεδομένων με πολύ ενθαρρυντικά αποτελέσματα. Πιο συγκεκριμένα τα αποτελέσματα επιδεικνύουν ότι οι παραγόμενοι ταξινομητές αποδίδουν καλύτερα από τον κλασσικό kNN και είναι πιο αξιόπιστοι χωρίς να είναι πιο αργοί.

Η kNN είναι μια μέθοδος που χρησιμοποιείται για αναγνώριση προτύπων και εφαρμόζεται σε ποικίλες περιπτώσεις (Monero-Seco et al.,2003; Abidin et al., 2006; Khan et al., 2002; Yu et al., 2002). Η απλότητα και η σχετικά υψηλή ταχύτητα σύγκλισης την καθιστούν μια πολύ διαδεδομένη επιλογή και ευρέως χρησιμοποιούμενη. Παρόλα αυτά σε μερικές εφαρμογές μπορεί να αποτύχει να παράγει επαρκή αποτελέσματα, ενώ σε άλλες μπορεί να μην είναι ουσιαστικά πρακτική (Sotoca et al., 2003; Monero-Seco et al., 2003). Το γεγονός βέβαια ότι έχει μόνο μία παράμετρο, τον αριθμό των γειτόνων που χρησιμοποιούνται (k), την καθιστά ικανή να ανταποκρίνεται σε μια ποικιλία καταστάσεων. Η κύρια διαδικασία της kNN τεχνικής αποτελείται από τα ακόλουθα βήματα: δεδομένου ενός συνόλου αποτελούμενο από *N* σημεία (το σύνολο εκπαίδευσης), του οποίου οι ετικέτες κλάσης είναι γνωστές, η kNN μέθοδος ταξινομεί ένα σύνολο από *n* σημεία (το σύνολο δοκιμής) μέσα στο ίδιο σύνολο κλάσεων εξετάζοντας τα *k* πλησιέστερα σημεία γύρω από κάθε σημείο του συνόλου δοκιμής και εφαρμόζοντας το πλειοψηφικό σχήμα-σύστημα ψηφοφορίας. Αυτή η διαδικασία έχει όμως αρκετά συμφυή προβλήματα και για το λόγο αυτό οι οι ερευνητές προσπάθησαν να τα λύσουν με διάφορες επεκτάσεις της kNN μεθόδου ή ολικούς μετασχηματισμούς των kNN ταξινομητών. Νέες εκδοχές, προσεγγίσεις της kNN τεχνικής που προτάθηκαν, παρουσιάζουν ενδιαφέροντα και πολλά υποσχόμενα αποτελέσματα, έχοντας σαν αποτέλεσμα την κινητοποίηση πολλών ερευνητών ώστε να προσπαθήσουν να βελτιώσουν τη kNN μέθοδο.

Αν και η kNN μέθοδος είναι αρκετά γρήγορη όταν χρησιμοποιείται σε προβλήματα ταξινόμησης, η όλη διαδικασία εμποδίζεται από το μέγεθος ορισμένων συνόλων δεδομένων. Για αυτό ακριβώς το λόγο κάποιοι ερευνητές προσπάθησαν να βελτιώσουν την ταχύτητα της μεθόδου. Ένα παράδειγμα είναι η SMART-TV (Abidin and Perizzo, 2006), η οποία σχεδιάστηκε για να ανταποκρίνεται σε σύνολα δεδομένων υψηλής διάστασης μετασχηματίζοντάς τα σε μονοδιάσταστο χώρο χαρακτηριστικών. Αυτές οι προσεγγίσεις επικεντρώνονται κυρίως μόνο στην υψηλή ταχύτητα και συχνά αποτυγχάνουν να επιτύχουν ένα πάρα πολύ καλό ποσοστό ακριβείας, εκτός και αν εφαρμόζονται σε συγκεκριμένα προβλήματα, όπως για

παράδειγμα στα χωρικά σύνολα δεδομένων (spatial data sets). Άλλες προσεγγίσεις της κNN μεθόδου που έχουν προταθεί μέχρι τώρα στην διεθνή βιβλιογραφία, περιλαμβάνουν μεθόδους επιλογής χαρακτηριστικών, χωρίς βέβαια να παρουσιάζουν θεαματικά αποτελέσματα όσον αφορά την βελτίωση της ταχύτητας. Μία παρόμοια μέθοδο, παραλλαγή του kNN θα παρουσιάσουμε διεξοδικά παρακάτω. Συχνά είναι πιο αποτελεσματικό να συνδυάζονται διαφορετικοί ταξινομητές με δύο διαφορετικούς τρόπους, είτε με την μορφοποίηση ενός χαμηλού επιπέδου μείγματος (Hendrickx and Antal van den Bosch, 2004), ή με την κατασκευή ενός συνόλου (Domeniconi and Yan, 2005). Στην πρώτη περίπτωση είναι εμφανές ότι οι αλλαγές στην δομή της kNN μεθόδου είναι σημαντικές για την βελτίωση της απόδοσης της μεθόδου. Αυτή η ιδέα προκάλεσε το ενδιαφέρον στους ερευνητές ώστε να αναπτύξουν νέους τύπους ταξινομητών βασισμένους στη kNN τεχνική. Στην δεύτερη περίπτωση, γίνεται προσπάθεια δημιουργίας ενός ασυσχέτιστου ταξινομητή, διότι οι αρνητικά συσχετισμένοι ταξινομητές μέσα ένα σύνολο φαίνεται να βελτιώνουν το ποσοστό ακριβείας του συνόλου. Τα αποτελέσματα αν και είναι ενδιαφέροντα, υποδηλώνουν ότι οι προσεγγίσεις που βασίζονται στη kNN τεχνική απαιτούν αρκετή βελτίωση αν είναι να χρησιμοποιηθούν σε σύνολα με στόχο τις διάφορες κλάσεις των προβλημάτων. Μία εναλλακτική μέθοδος η οποία έχει ερευνηθεί είναι η χρήση κανόνων στη kNN μέθοδο, στην οποία οι κανόνες έχουν χρησιμοποιηθεί ως πρόσθετα γνωρίσματα-χαρακτηριστικά με μερική επιτυχία (Antal van den Bosch, 2004). Παρόλα αυτά σε μερικά σύνολα δεδομένων η δημιουργία κανόνων μπορεί να είναι χρονοβόρα και να κοστίζει υπολογιστικά. Επίσης, σε σύνολα δεδομένων υψηλής διάστασης το πρόσθετο κόστος μπορεί να καταστήσει την ταξινόμηση πολύ αργή και συνεπώς ανεπαρκή.

Μια άλλη μέθοδος εξετάζει την αποτελεσματικότητα του συστήματος-σχήματος ψηφοφορίας της kNN μεθόδου, και προτείνει ένα εναλλακτικό μέτρο που στόχο έχει να καθορίσει πως κάθε τάξη σχετίζεται με ένα σημείο δοκιμής (Wang and Bell, 2004). Αυτή η μέθοδος πέρα από την απλή καταμέτρηση των γειτόνων ταυτόχρονα επιτυγχάνει και την αξιολόγησή τους.

Βασισμένοι σε αυτή τη φιλοσοφία, οι Zacharias Voulgaris και George D. Magoulas παρουσίασαν εναλλακτικές τεχνικές της kNN μεθόδου που είτε αντιστοιχούν έναν ποιοτικό δείκτη σε κάθε στοιχείο του συνόλου δεδομένων, ή μια διαφορετική *k* τιμή. Επίσης, εισήγαγαν έναν ταξινομητή με παρόμοιο τρόπο όπως στην μέθοδο που περιγράφεται στην εργασία των Sotoca et al. (2003), όπου χρησιμοποιούνται διαφορετικά βάρη για τα διάφορα χαρακτηριστικά του συνόλου δεδομένων και μάλιστα αυτό επιτυγχάνεται με έναν γρήγορο και αρκετά αποδοτικό τρόπο.
Δύο μέτρα είναι αυτά τα οποία χρησιμοποιούνται ευρέως για την εκτίμηση της απόδοσης των παραγόμενων ταξινομητών και επιπρόσθετα εκτιμούν με διορατικότητα και την απόδοση των μεθόδων με μετασχηματισμούς συνόλου. Το πρώτο μέτρο είναι ο βαθμός βεβαιότητας, το οποίο μας δείχνει το κατά πόσο βέβαιοι είναι οι ταξινομητές για κάθε ταξινόμηση που πραγματοποιείται. Το δεύτερο μέτρο

είναι το δίκτυο αξιοπιστίας, το οποίο μετρά το κατά πόσο σχετίζεται η βεβαιότητα της ταξινόμησης με την ορθότητά της.

- Βαθμός Βεβαιότητας

Ο Βαθμός βεβαιότητας είναι μια γενίκευση του συντελεστή βεβαιότητας (Certainty Factor, CF) που ορίζεται για έναν τύπο ταξινομητή ως:

$$CF_i = \frac{final\ vote(i)}{\sum_{c=1}^{\#of\ classes} final\ vote(c)} \quad (1)$$

όπου το i δηλώνει το i-οστό ταξινομημένο μοτίβο και το c αντιστοιχεί στον αριθμό της κλάσης.
Αντικαθιστώντας το $final\ vote(c)$ με την βαθμό ταξινόμησης της κλάσης και το $final\ vote(i)$ με το $max(final\ vote)$, προκύπτει ο βαθμός βεβαιότητας (Degree of Certainty, CF), ένα δείκτη βεβαιότητας που είναι συμβατός με όλους τους τύπους ταξινομητών. Αυτό το μέτρο αποφέρει πληροφορία για το βαθμό εμπιστοσύνης του ταξινομητή για μια συγκεκριμένη ταξινόμηση και έχει την μορφή διανύσματος:

$$DC_i = \frac{\max(classification\ score)}{\sum_{c=1}^{\#of\ classes} classification\ score(c)} \quad (2)$$

όπου το i δηλώνει το i-οστό ταξινομημένο μοτίβο-πρότυπο (pattern), c ο αριθμός της κλάσης και $classification\ score$ είναι το score το οποίο καθορίζει το αποτέλεσμα της ταξινόμησης που πρόεκυψε από έναν ταξινομητή.

- Δίκτυο Αξιοπιστίας

Το δίκτυο αξιοπιστίας (Net Reliability, NR) είναι ένα μέτρο το οποίο πρωταρχικά σχεδιάστηκε με στόχο να εκτιμήσει το πόσο αξιόπιστος είναι ο βαθμός βεβαιότητας (DC) ενός ταξινομητή. Είχε παρατηρηθεί ότι υπάρχουν περιπτώσεις όπου ένας ταξινομητής έχει υψηλό βαθμό βεβαιότητας για μια ταξινόμηση, η οποία αργότερα αποδεικνυόταν λάθος, ενώ για χαμηλό βαθμό βεβαιότητας είχαμε σωστή ταξινόμηση. Με άλλα λόγια, το μέτρο αυτό είναι παρόμοιο με ένα μέτρο συσχέτισης μεταξύ Ακριβείας και Βαθμού βεβαιότητας, που παίρνει τιμές στο διάστημα [-1,1]. Είναι ολοφάνερο ότι όσο υψηλότερο είναι το δίκτυο αξιοπιστίας του ταξινομητή, τόσο το καλύτερο για τον ταξινομητή (συνήθως οτιδήποτε θετικό είναι και καλό). Το δίκτυο αξιοπιστίας ορίζεται από την παρακάτω σχέση:

$$NR = \frac{1}{n}\sum_{i=1}^{n}[(2v_i - 1) \cdot DCy_i] \quad (3)$$

όπου v είναι το διάνυσμα ισχύος ταξινόμησης, δηλαδή το διάνυσμα της εγκυρότητας μιας ταξινόμησης (ένα δυαδικό διάνυσμα, που απεικονίζει τις σωστές ταξινομήσεις με 1 και τις λάθος με 0), DCy το διάνυσμα του βαθμού βεβαιότητας, i το ταξινομημένο μοτίβο, n ο συνολικός αριθμός στοιχείων στο σύνολο δοκιμής.
Επειδή το NR εξαρτάται από τον ταξινομητή καθώς και από το σύνολο δεδομένων, είναι χρήσιμο να υπολογίζεται κάθε φορά και να θεωρείται σημαντικό μέτρο της απόδοσής του μέτρου αβεβαιότητας, αν χρησιμοποιείται το μέτρο αβεβαιότητας του ταξινομητή.

2.2 Εκδοχές Ταξινόμησης που βασίζονται στο kNN

2.2.1 Ο kNN ταξινομητής που βασίζεται στην πυκνότητα The Density Based kNN classifier, (DB-kNN)

Από την στιγμή που η απλή καταμέτρηση των γειτόνων φαίνεται να είναι ανεπαρκής για τον καθορισμό της κλάσης ενός στοιχείου ελέγχου, οι Zacharias Voulgaris και George D. Magoulas τροποποίησαν τον kNN ταξινομητή θεωρώντας σημαντικό ένα άλλο παράγοντα, την πυκνότητα, η οποία καλείται Διαρθρωτική Πυκνότητα (Structural Density, SD), καθώς παρέχει πληροφορία για την δομή ενός συνόλου δεδομένων και είναι εμπνευσμένη από τις Φυσικές Επιστήμες. Η διαρθρωτική πυκνότητα ορίζεται ως ο αριθμός των σημείων στην περιοχή-γειτονιά ενός στοιχείου επί τον όγκο αυτής της περιοχής. Η παράμετρος που περιλαμβάνεται είναι η ακτίνα (r), που ορίζει την περιοχή, η οποία καθορίζεται από τα ακόλουθα βήματα:

- Υπολογισμός της πυκνότητας όλων των στοιχείων ως συνάρτηση του r και της μέσης πυκνότητας όλου του συνόλου, ως ο συνολικός αριθμός στοιχείων επί τον συνολικό όγκο του συνόλου δεδομένων, που δεν σχετίζεται με το r.
- Εύρεση της τιμής του r για την οποία η μέση τιμή των μεμονωμένων πυκνοτήτων να ισούται με την μέση πυκνότητα που υπολογίζεται νωρίτερα.

Αρχικά οι πυκνότητες όλων των στοιχείων υπολογίζονται για καθεμιά από τις κλάσεις του συνόλου δεδομένων. Στην συνέχεια, γίνεται κανονικοποίηση στο διάστημα [0,1] εφόσον οι σχετικές πυκνότητες φαίνεται να είναι πιο χρήσιμες από τις απόλυτες συχνότητες. Βασισμένοι σε τέτοιες πυκνότητες, κάθε γειτονικό στοιχείο αξιολογείται όσον αφορά το ρόλο του σαν στοιχείο του πυρήνα της κλάσης του, μετρώντας την διαρθρωτική σχετική πυκνότητα όσον αφορά την κλάση. Διαιρώντας το με την Ευκλείδια απόσταση, παρέχεται ένα αποτέλεσμα (score) για κάθε γειτονικό στοιχείο. Θεωρώντας ένα μεροληπτικό μέσο αυτών των αποτελεσμάτων (scores) για κάθε κλάση καταλήγουμε με q αποτελέσματα ψηφοφορίας, όπου q είναι ο αριθμός των κλάσεων. Αυτός ο μέσος επηρεάζεται περισσότερο από μεγαλύτερους αριθμούς. Από αυτά τα τελικά scores προκύπτει η ταξινόμηση κάθε στοιχείου ελέγχου, καθώς και ο βαθμός βεβαιότητας (DC). Σε περιπτώσεις (που βέβαια δεν συναντάμε συχνά)

όπου ο βαθμός βεβαιότητας της ταξινόμησης ενός συγκεκριμένου στοιχείου ελέγχου είναι μικρότερος από 0.667, εφαρμόζεται η κλασσική μέθοδος kNN.
Η μέθοδος DB-kNN προσφέρει μια νέα χρήση των γειτόνων στην kNN επειδή εξερευνά την δυνατότητα αξιολόγησής τους αντί για την απλή καταμέτρησή τους. Επίσης το γεγονός ότι χρησιμοποιείται η απόσταση που καθιστά της αξιολόγηση των γειτόνων πιο εξευγενισμένη.

2.2.2 kNN ταξινομητής μεταβλητής (The variable k nearest neighbor classifier-V-kNN)

Από την στιγμή που η η τιμή της παραμέτρου *k* συχνά επηρεάζει τα αποτελέσματα της ταξινόμησης, και κάποιες φορές πολύ σημαντικά, αναπτύχθηκε ένας άλλος αλγόριθμος ταξινόμησης που ξεπερνά το πρόβλημα αυτό. Κάνοντας χρήση του DC υπολογίζεται το βέλτιστο *k* για κάθε ταξινόμηση.
Ο ταξινομητής V-kNN λειτουργεί ως εξής:

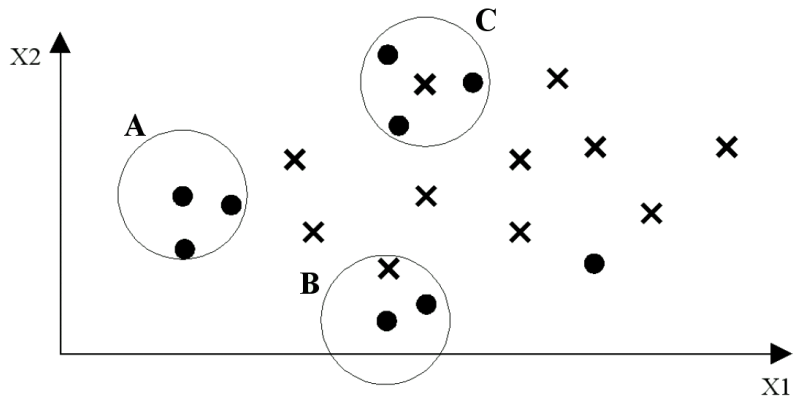
- Για κάθε ένα από τα στοιχεία του συνόλου εκπαίδευσης γίνεται μια ταξινόμηση η οποία βασίζεται σε διάφορους γείτονες.
- Βρίσκεται η *k* τιμή που μεγιστοποιεί τον βαθμό βεβαιότητας της κάθε ταξινόμησης.
- Σε κάθε εκπαιδευτικό σύνολο αντιστοιχεί μια συγκεκριμένη τιμή *k* η οποία θεωρείται η καλύτερη διαθέσιμη.
- Έπειτα για κάθε άγνωστο στοιχείο βρίσκεται ο πλησιέστερος γείτονας και ο πειραματιστής υποθέτει την *k* τιμή του, βασισμένος στη βέλτιστη *k* συστοιχία.
- Ο ταξινομητής kNN εφαρμόζεται σε αυτό το στοιχείο, χρησιμοποιώντας την υποτιθέμενη τιμή του *k.*

Η προσέγγιση V-kNN αποφέρει χρήσιμες πληροφορίες για το σύνολο δεδομένων, πέρα από την απόδοσή της σαν ταξινομητής. Η προσέγγιση V-kNN εκτιμά την μέση βέλτιστη τιμή του *k,* η οποία για τους ταξινομητές τύπου kNN είναι πολύ χρήσιμη καθώς βελτιώνει την απόδοσή τους, ειδικά όταν πρόκειται για τον κλασσικό kNN ταξινομητή. Παρόλα αυτά, για ένα σποραδικό-αραιό σύνολο δεδομένων, η βέλτιστη τιμή του *k* μπορεί να μην είναι έγκυρη και τα αποτελέσματα μπορεί να μην είναι καλύτερα από αυτά του kNN.

2.2.3 Ο σταθμισμένος kNN ταξινομητής (The weighted kNN classifier) W-kNN

Ο σταθμισμένος kNN ταξινομητής εκτελεί μια αξιολόγηση, με παρόμοιο τρόπο με αυτό του kNN ταξινομητή που βασίζεται στην πυκνότητα, αυτή τη φορά όμως η αξιολόγηση γίνεται στα χαρακτηριστικά αντί στα πρότυπα. Κάθε χαρακτηριστικό αξιολογείται και του εκχωρείται μια μεταβλητή βάρους, η οποία βασίζεται στην χρησιμότητα του χαρακτηριστικού για την διάκριση των κλάσεων του συνόλου δεδομένων. Για να είναι κάτι τέτοιο εφικτό μια νέα έννοια εισήχθη (Domeniconi and Yan, 2005; Wang and Bell, 2004): Ο Δείκτης διακριτικότητας (Index of Discernibility-ID).

Ο δείκτης διακριτικότητας είναι ένα μέτρο που αναπτύχθηκε για την αξιολόγηση της ευκολίας διάκρισης των κλάσεων σε ένα σύνολο δεδομένων. Αρχικά διακρίνονται οι κλάσεις με την χρήση boxes που περιείχαν τις κλάσεις αυτές, αλλά αυτή η τεχνική δεν ήταν ευαίσθητη στην δομή της κλάσης ούτε υπολογιστικά αποτελεσματική. Για αυτή την εκδοχή του δείκτη διακριτικότητας, χρησιμοποιούνται σφαίρες - (hyper)spheres με σταθερή ακτίνα γύρω από κάθε στοιχείο του συνόλου δεδομένων, η οποία αντιστοιχεί στη μέση απόσταση μεταξύ του στοιχείου αυτού και των υπόλοιπων στοιχείων της κλάσης. Η ακτίνα αυτή βασίζεται στην δομή της κλάσης και για αυτό το λόγο τα στοιχεία που ανήκουν σε διαφορετικές κλάσεις είναι πιθανό να έχουν διαφορετικές ακτίνες. Όταν καθορίζεται η ακτίνα ενός στοιχείου, προσδιορίζονται και καταμετρώνται τα στοιχεία της ίδιας κλάσης με το υπό εξέταση στοιχείο που ανήκουν στην σφαίρα. Η διακριτικότητα του στοιχείου υπολογίζεται διαιρώντας τον αριθμό αυτών των στοιχείων με τον συνολικό αριθμό των στοιχείων που βρίσκονται στην σφαίρα, όπως απεικονίζεται από το Σχήμα 1. Ο δείκτης Διακριτικότητας όλου του συνόλου δεδομένων υπολογίζεται ως ο αριθμός των στοιχείων που έχουν διακριτικότητα μεγαλύτερη από 0.5 διαιρούμενος με τον συνολικό αριθμό στοιχείων.
Ο δείκτης διακριτικότητας χρησιμοποιείται επίσης για την αξιολόγηση ξεχωριστών χαρακτηριστικών απλά με την μονοδιάστατη εφαρμογή του στο σύνολο δεδομένων.



Σχήμα 1- Εικόνα του Δείκτη Διακριτικότητας-ID για ένα απλό σύνολο δεδομένων με δύο μοναδικά στοιχεία, X1 και X2. Σε αυτό το παράδειγμα, η διακριτικότητα του στοιχείου στο κέντρο του κύκλου A είναι ID1=2/2=1, η διακριτικότητα του στοιχείου στο κέντρο του κύκλου B είναι ID2=1/2=0.5, και η διακριτικότητα του στοιχείου στο κέντρο του κύκλου C είναι ID3=0/3=0.

Ο ταξινομητής W-kNN λειτουργεί ως εξής:

- Κάθε ένα στοιχείο του συνόλου εκπαίδευσης αξιολογείται χρησιμοποιώντας τον ID.
- Έπειτα προκύπτουν τα βάρη κανονικοποιώντας τα IDs.
- Τελικά, τα βάρη εφαρμόζονται στο σύνολο εκπαίδευσης και στο σύνολο δοκιμής και ο ταξινομητής kNN εφαρμόζεται πλέον στο νέο αυτό μετασχηματισμένο σύνολο δεδομένων.

2.2.4 Ο ταξινομητής kNN που βασίζεται στην κλάση (The class based kNN classifier- CB-kNN)

Ο ταξινομητής CB-kNN δημιουργήθηκε επειδή σε μερικές περιπτώσεις τα σύνολα δεδομένων δεν είναι ισορροπημένα όσον αφορά την δομή της κλάσης τους. Είναι πιθανό λοιπόν σε μια τέτοια περίπτωση μια κλάση να έχει πολύ λίγα στοιχεία ώστε να κερδίσει την ψήφο της ταξινόμησης του ταξινομητή kNN.
Ο αλγόριθμος CB-kNN λειτουργεί ως εξής:

- Για κάθε στοιχείο ελέγχου, επιλέγονται τα k πλησιέστερα στοιχεία κάθε κλάσης.
- Η τιμή του k επιλέγεται αυτόματα από τον ταξινομητή, ώστε να μεγιστοποιείται ο DC της ταξινόμησης.
- Έπειτα, υπολογίζεται ο αρμονικός μέσος των αποστάσεων αυτών των γειτόνων, έτσι ώστε να μην επηρεάζεται από τα πιο απομακρυσμένα στοιχεία.
- Στο τελικό στάδιο, αυτοί οι αρμονικοί μέσοι συγκρίνονται και για την ταξινόμηση επιλέγεται η κλάση που εμφανίζεται να έχει την μικρότερη τιμή.

2.2.5 Ο ταξινομητής kNN διάκρισης (The discernibility kNN classifier - D-kNN)

Εμπνευσμένοι από την μέθοδο W-kNN, οι Zacharias Voulgaris και George D. Magoulas χρησιμοποίησαν ξανά την έννοια της διάκρισης με έναν διαφορετικό αυτή τη φορά τρόπο. Ο αλγόριθμος έχει παρόμοια δομή με την αρχική kNN επέκταση του DB-kNN, με στόχο να κατασκευαστεί ένας αλγόριθμος πολύ γρήγορος και ταυτόχρονα ακριβής. Ο D-kNN λαμβάνει υπόψη του την απόσταση κάθε γείτονα, παρόμοια με τον DB-kNN, μόνο που αντί για την πυκνότητα της δομής, χρησιμοποιεί την ικανότητα διάκρισης κάθε στοιχείου. Διαιρώντας την διακριτικότητα με την απόσταση, προκύπτει ένα score για κάθε έναν από τους γείτονες. Έπειτα, υπολογίζεται ο μέσος όρος των scores για κάθε κλάση και προκύπτει με αυτό το τρόπο ένα score ταξινόμησης για κάθε μία από τις κλάσεις. Η κλάση που φέρει το υψηλότερο score ταξινόμησης επιλέγεται για την ταξινόμηση.

2.3 Πειραματικά Αποτελέσματα

2.3.1 Περιγραφή πειραμάτων

Τα πειράματα που διενεργήθηκαν περιελάμβαναν 500 ταξινομήσεις (50 γύρους 10-fold cross validation) για κάθε έναν ταξινομητή. Εκτελέστηκαν σε 6 σύνολα δεδομένων προερχόμενα από την UCI (UCI Machine Learning Repository). Τα χαρακτηριστικά αυτών των συνόλων δεδομένων περιγράφονται συνοπτικά στον Πίνακα 1. Οι ταξινομητές που απαιτούν μια τιμή k χρησιμοποίησαν αυτήν που βρίσκεται στην δεξιότερη στήλη του πίνακα.

Dataset Characteristics			
Name	**Attributes**	**Patterns**	**K***
Bupa Liver	6	345	7
Pima Indians	8	768	6
Breast Cancer W.	9	683	2
Heart Disease	13	270	6
Vehicle	18	846	5
Boston Housing	13	506	6

Πίνακας 1.
Τα χαρακτηριστικά των συνόλων δεδομένων που χρησιμοποιούνται στα πειράματα. Η στήλη K* δείχνει το βέλτιστο *k* που υπολογίστηκε από τον Variable *k* Nearest Neighbor αλγόριθμο.

2.3.2 Αποτελέσματα

Η απόδοση των ταξινομητών οι οποίοι βασίζονται στην kNN τεχνική αξιολογήθηκε με την χρήση του ποσοστού ακριβείας (AR) και του δικτύου αξιοπιστίας (NR) που παρουσιάστηκαν αναλυτικά προηγουμένως, αλλά και με την χρήση του μέσου CPU χρόνου του εκπαιδευτικού μέρους της ταξινόμησης. Για κάθε ένα από τα σύνολα δεδομένων, πραγματοποιήθηκε μια σειρά πειραμάτων και με βάση τα παραπάνω κριτήρια βρέθηκε ο «νικητής», δηλαδή ο καλύτερος διαθέσιμος ταξινομητής. Τα αποτελέσματα απεικονίζονται στον Πίνακα 2.

Dataset	**Accuracy Rate**	**Net Reliability**	**CPU Time (2nd) in sec.**
Bupa Liver	D-kNN (66.31%)	D-kNN (0.2572)	W-kNN (0.0226)
Pima Indians	V-kNN (74.55%)	V-kNN (0.4707)	W-kNN (0.1020)
Breast Cancer W.	CB-kNN (96.93%)	V-kNN (0.9285)	W-kNN (0.0920)
Heart Disease	D-kNN (81.31%)	D-kNN (0.5926)	W-kNN (0.0332)
Vehicle	CB-kNN (70.89%)	D-kNN (0.4009)	V-kNN (0.1230)
Boston Housing	CB-kNN (69.03%)	W-kNN (0.3587)	W-kNN (0.0789)

Πίνακας 2. - Οι νικητές με βάση την μέση απόδοση στους 50 γύρους. Το ποσοστό της απόδοσης νίκης φαίνεται στις παρενθέσεις. Το Δίκτυο Αξιοπιστίας υπολογίζεται μέσω της ταξινόμησης και των διανυσμάτων του Βαθμού Βεβαιότητας στο τέλος κάθε πειράματος.

Στην συνέχεια, έγινε μια ένα-προς-ένα σύγκριση για κάθε ζευγάρι ταξινομητών, η οποία δείχνει πόσες φορές, δηλαδή πόσους γύρους ο ένας ταξινομητής υπερτερεί του άλλου. Έπειτα, τα scores αυτά αθροίζονται για κάθε ταξινομητή. Το τελικό άθροισμα αποκαλύπτει την σχετική απόδοση κάθε ταξινομητή, όπως φαίνεται στον Πίνακα 3.

Dataset	Accuracy Rate	Net Reliability	CPU Time (2nd) in sec.
Bupa Liver	D-kNN (232)	D-kNN (241)	W-kNN (200)
Pima Indians	V-kNN (219)	V-kNN (250)	V-kNN (193)
Breast Cancer W.	V-kNN (230)	V-kNN (237)	CB-kNN (200)
Heart Disease	D-kNN (193)	D-kNN (226)	V-kNN (200)
Vehicle	CB-kNN (218)	D-kNN (231)	V-kNN (200)
Boston Housing	W-kNN (210) & CB-kNN (209)	W-kNN (214)	V-kNN (200)

Πίνακας 3. - Οι νικητές με βάση την σχετική απόδοση, σε ζευγάρια, πάνω από 50 γύρους. Οι αριθμοί στις παρενθέσεις δείχνουν το άθροισμα των φορών που ο νικητήριος ταξινομητής ήταν καλύτερος από τους άλλους, σύμφωνα με ένα συγκεκριμένο μέτρο απόδοσης, για κάθε σύνολο δεδομένων.

Είναι αξιοσημείωτο ότι και για τα έξι σύνολα δεδομένων, βρέθηκε τουλάχιστον μία παραλλαγή της kNN τεχνικής από όλες αυτές που παρουσιάστηκαν πριν αναλυτικά, να υπερτερεί του kNN. Το μοναδικό κριτήριο μάλιστα στο οποίο αποδίδει καλά ο kNN είναι η ταχύτητα, διότι δεν χρειάζεται εκπαίδευση.

2.4 Συμπεράσματα

Από την παραπάνω λεπτομερή μελέτη στις διάφορες πιθανές επεκτάσεις της μεθόδου kNN, η προσέγγιση που χρησιμοποιεί ιδιότητες του συνόλου δεδομένων οδήγησε στην ανάπτυξη αποτελεσματικών μετασχηματισμών της μεθόδου kNN για τα προβλήματα ταξινόμησης. Οι προτεινόμενες kNN παραλλαγές εξετάστηκαν ως προς την απόδοση τους σε προβλήματα ταξινόμησης χρησιμοποιώντας δεδομένα προερχόμενα από την UCI.

Με βάση τα πειράματα που πραγματοποιήθηκαν η μέθοδος DB-kNN είναι πιο αργή από την kNN εξαιτίας του υπολογισμού της πυκνότητας της δομής, αλλά γενικά έχει καλύτερη απόδοση από την kNN.

Ο ταξινομητής V-kNN είναι πολύ γρήγορος, ίσως ο γρηγορότερος σε σχέση με όλους τους άλλους, και γενικά έχει καλύτερες αποδόσεις από τον kNN.

Η W-kNN μέθοδος είναι εξαιρετικά γρήγορη, διότι οι λειτουργίες που χρειάζονται για τον υπολογισμό και την εφαρμογή των μεταβλητών βάρους είναι πολύ απλές. Η απόδοσή της είναι καλύτερη από της kNN και ο χρόνος CPU που χρειάζεται είναι ο ελάχιστος για διάφορες τιμές του *k*.

Όσον αφορά τον ταξινομητή CB-kNN, ο μεγάλος αριθμός των υπολογισμών που περιλαμβάνονται στην διαδικασία υλοποίησης του, καθιστά τον συνολικό CPU χρόνο μεγαλύτερο από αυτόν των άλλων ταξινομητών. Επίσης, υπερτερεί του kNN σημαντικά καθώς είναι ο πιο ακριβής ταξινομητής σε τρία από τα σύνολα δεδομένων.

Ο ταξινομητής D-kNN είναι ικανοποιητικά γρήγορος (όχι ο γρηγορότερος βέβαια), ενώ ταυτόχρονα υπερτερεί του kNN και κάποιων επεκτάσεών αυτού. Επίσης έχει αποδειχθεί αρκετά αξιόπιστος από την άποψη του Δικτύου Αξιοπιστίας.

Ανακεφαλαιώνοντας όλοι οι ταξινομητές που παρουσιάστηκαν και εξετάστηκαν έχουν καλή απόδοση, αν και κάποιοι ξεχωρίζουν σε μερικά σύνολα δεδομένων. Οι ταξινομητές αυτοί αν και είναι πιο αργοί από τον kNN, επειδή χρειάζονται μια φάση εκπαίδευσης, είναι όλοι αξιόπιστοι με την έννοια του Δικτύου Αξιοπιστίας (NR). Χρειάζεται βέβαια περαιτέρω έλεγχος σε μεγαλύτερη κλίμακα ώστε να διερευνηθούν πλήρως τα πλεονεκτήματα των προτεινόμενων μεθόδων αυτών καθώς και οι περιορισμοί τους.

3 ΚΕΦΑΛΑΙΟ – Η ΜΕΘΟΔΟΣ SVM

3.1 Εισαγωγή στις Μηχανές υποστήριξης διανυσμάτων (Support Vector Machines, SVM)

Η εκμάθηση μηχανών θεωρείται ένα από τα πεδία της τεχνητής νοημοσύνης και ασχολείται με την ανάπτυξη τεχνικών και μεθόδων που καθιστούν τον υπολογιστή ικανό να μάθει και να εκτελεί εργασίες. Η εκμάθηση μηχανών έχει κάποια κοινά χαρακτηριστικά με την στατιστική. Με το πέρασμα του χρόνου αναπτύχθηκαν πολλές τεχνικές και μεθοδολογίες για την εκπαίδευση των μηχανών να εκτελούν εργασίες.

Ο όρος Support Vector Machine (Μηχανή υποστήριξης διανύσματος) εμφανίστηκε το 1992, και εισήχθη από τους Boser, Guyon και Vapnik στο COLT-92. Οι Μηχανές υποστήριξης διανυσμάτων είναι ένα σύνολο συσχετισμένων μεθόδων εκμάθησης με επίβλεψη, που χρησιμοποιούνται στην ταξινόμηση και την παλινδρόμηση και ανήκουν σε μια οικογένεια γενικευμένων γραμμικών ταξινομητών. Αποτελούν ένα εργαλείο πρόβλεψης για την ταξινόμηση και την παλινδρόμηση, που χρησιμοποιεί θεωρία εκμάθησης μηχανών ώστε να μεγιστοποιήσει την ακρίβεια πρόβλεψης ενώ αυτόματα αποφεύγει το overfitting στα δεδομένα. Οι Μηχανές υποστήριξης διανυσμάτων (SVMs) μπορούν να οριστούν ως συστήματα που χρησιμοποιούν υποθετικό χώρο γραμμικών συναρτήσεων σε έναν υψηλής διάστασης χώρο χαρακτηριστικών, τα οποία έχουν εκπαιδευτεί με έναν αλγόριθμο εκμάθησης προερχόμενο από την θεωρία βελτιστοποίησης, ο οποίος ενσωματώνει τη μεροληψία (bias) εκμάθησης η οποία προέρχεται από την στατιστική θεωρία εκμάθησης. Αρχικά, οι Μηχανές υποστήριξης διανυσμάτων ήταν ευρέως γνωστές στην NIPS κοινότητα και τώρα αποτελούν ενεργό μέρος της έρευνας των μηχανών εκμάθησης.

Οι Μηχανές υποστήριξης διανυσμάτων έγιναν γνωστές όταν κατά την χρήση pixel maps ως δεδομένα εισόδου, παρείχαν αντίστοιχη ακρίβεια με αυτή των εξελιγμένων νευρωνικών δικτύων για την επεξεργασία χαρακτηριστικών σε μια εργασία αναγνώρισης γραφικού χαρακτήρα του A.W. Moore, (2003). Φυσικά, οι Μηχανές υποστήριξης διανυσμάτων χρησιμοποιούνται σε πολλές ακόμα εφαρμογές, όπως για παράδειγμα η ανάλυση του γραφικού χαρακτήρα, ανάλυση προσώπου και ειδικά σε εφαρμογές που βασίζονται σε ταξινόμηση και παλινδρόμηση προτύπων. Τα θεμέλια των SVMs έχουν αναπτυχθεί από τον Vapnik (Vapnik, 1995), και κέρδισαν φήμη εξαιτίας των πολλά υποσχόμενων χαρακτηριστικών, όπως είναι η καλύτερη εμπειρική επίδοση. Στις Μηχανές υποστήριξης διανυσμάτων χρησιμοποιείται η αρχή της Structural Risk Minimization (SRM) (Burges B., 1998), που έχει αποδειχθεί ότι είναι καλύτερη σε σχέση με την παραδοσιακή αρχή της ελαχιστοποίησης του εμπειρικού ρίσκου (Empirical Risk Minimization-ERM) η οποία χρησιμοποιείται από τα συμβατικά νευρωνικά δίκτυα. Η SRM ελαχιστοποιεί ένα άνω φράγμα στο αναμενόμενο ρίσκο, ενώ η ERM ελαχιστοποιεί το σφάλμα στα δεδομένα εκπαίδευσης. Αυτή η διαφορά επιτρέπει στην SVM μέθοδο να έχει μεγαλύτερη δυνατότητα γενίκευσης που αποτελεί τον επιδιωκόμενο στόχο στη στατιστική θεωρία

εκμάθησης. Οι Μηχανές υποστήριξης διανυσμάτων κυρίως αναπτύχθηκαν για να λύνουν προβλήματα ταξινόμησης, αλλά επεκτάθηκαν και στην επίλυση προβλημάτων παλινδρόμησης (Vapnik et al. 1997).

3.2 Στατιστική θεωρία εκμάθησης και Μηχανές Εκμάθησης

Η στατιστική θεωρία εκμάθησης παρέχει ένα πλαίσιο μελέτης του προβλήματος απόκτησης γνώσης, πρόβλεψης, λήψης αποφάσεων από ένα σύνολο δεδομένων. Με απλά λόγια, καθιστά δυνατή την επιλογή του κατάλληλου υπερεπιπέδου με τέτοιο τρόπο ώστε να απεικονίζει την συνάρτηση ενδιαφέροντος στον χώρο στόχο (Evgeniou and Pontil, 1998).

Στην στατιστική θεωρία εκμάθησης το πρόβλημα της μάθησης με επίβλεψη διατυπώνεται ως εξής: Δίνεται το σύνολο των δεδομένων εκπαίδευσης $\{(x_1, y_1), \dots, (x_l, y_l)\}$ στον χώρο $R^n \times R$ που ακολουθεί μια άγνωστη κατανομή πιθανότητας $P(x, y)$, και μια συνάρτηση απώλειας $V(y, f(x))$ που μετράει το σφάλμα. Για ένα δοθέν x, προβλέπεται η τιμή $f(x)$ αντί για την πραγματική τιμή y. Το πρόβλημα συνίσταται στην εύρεση μιας συνάρτησης f που ελαχιστοποιεί την πρόβλεψη του σφάλματος στα νέα δεδομένα, δηλαδή μια συνάρτηση που να ελαχιστοποιεί το αναμενόμενο σφάλμα (Evgeniou and Pontil, 1998)

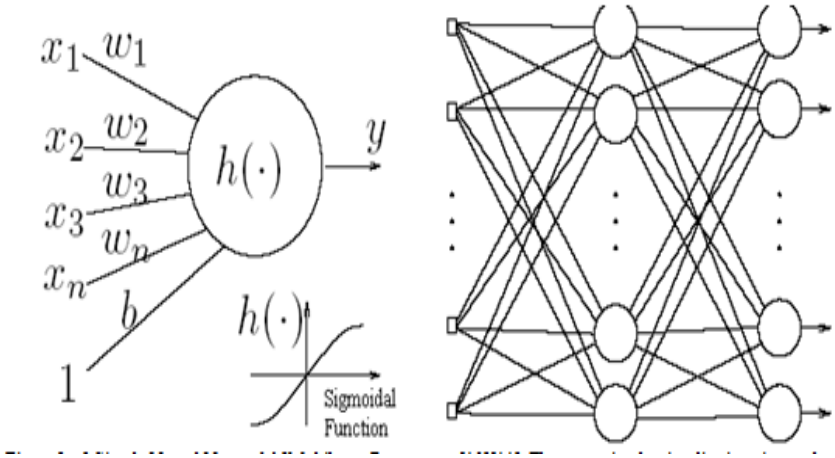
$$\int V(y, f(x))P(x, y)dxdy.$$

Στην στατιστική μοντελοποίηση θα διαλέγαμε ένα μοντέλο από τον χώρο της υπόθεσης, που είναι πλησιέστερο στην βασική συνάρτηση στον χώρο στόχο.
Οι πρώτοι αλγόριθμοι μηχανών εκμάθησης στόχευαν στην εκμάθηση αναπαράστασης απλών συναρτήσεων. Έτσι, ο στόχος της εκμάθησης ήταν να εξάγει μια υπόθεση που απέδιδε την σωστή ταξινόμηση των δεδομένων εκπαίδευσης και οι πρώτοι αλγόριθμοι σχεδιάστηκαν για την εύρεση ενός ακριβούς ταιριάσματος στα δεδομένα (Cristianini and Shawe-Taylor, 2000). Η ικανότητα μιας υπόθεσης να ταξινομεί σωστά τα δεδομένα όχι στο σύνολο εκπαίδευσης αποτελεί τη γενίκευση της μεθόδου. Η SVM τεχνική αποδίδει καλύτερα σε συνθήκες μη υπεργενίκευσης όταν τα νευρωνικά δίκτυα καταλήγουν εύκολα σε υπεργενίκευση (Mitchell, 1997). Ένα σημαντικό πρόβλημα, που πρέπει να αναφέρουμε, είναι η εύρεση του σημείου που γίνεται η καλύτερη εξισορρόπηση των παραγόντων στην πολυπλοκότητα του μοντέλου σε σχέση με τον αριθμό των εποχών. Στην ακόλουθη εικόνα απεικονίζεται η πολυπλοκότητα σε σχέση με τον αριθμό των εποχών.

Εικόνα 1. - Διάγραμμα του αριθμού των εποχών σε σχέση με την πολυπλοκότητα.

3.3 Μηχανές υποστήριξης διανυσμάτων (Support Vector Machines, SVM)

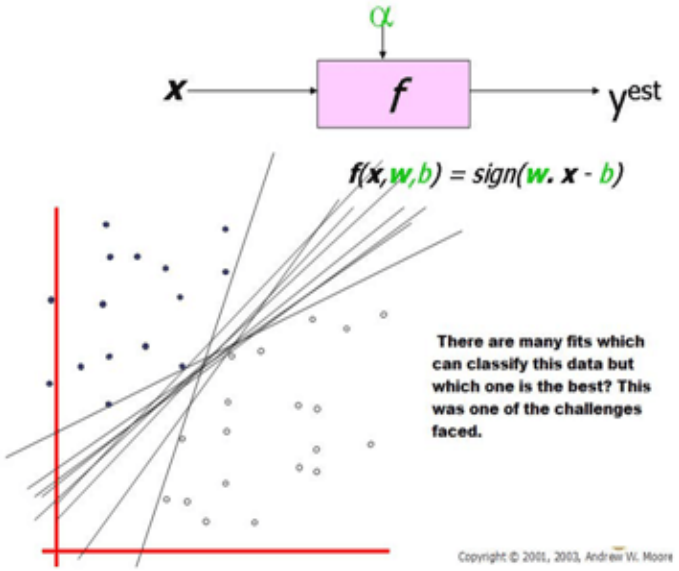
Η αρχική ενασχόληση με τα νευρωνικά δίκτυα για εκμάθηση με επίβλεψη και χωρίς επίβλεψη, παρουσίασε καλά αποτελέσματα ενώ χρησιμοποιήθηκε για ποικίλες εφαρμογές εκμάθησης. Ο Multilayer Perceptron (MLP) νευρώνας χρησιμοποιεί δίκτυα με προς τα εμπρός τροφοδότηση και επαναλαμβανόμενα δίκτυα. Οι ιδιότητες του MLP περιλαμβάνουν ολική προσέγγιση των συνεχών μη γραμμικών συναρτήσεων και προηγμένες αρχιτεκτονικές δικτύων με πολλαπλές εισόδους και εξόδους (Scapura, 1996).



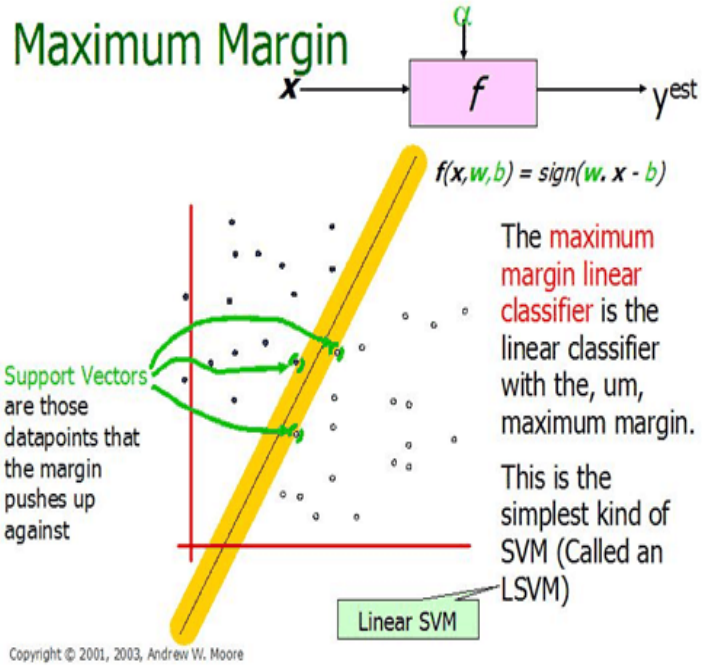
Εικόνα 2. - Απλό νευρωνικό δίκτυο και ο Multilayer Perceptron.

Διάφορα θέματα μπορεί να προκύψουν στην εφαρμογή των νευρωνικών δικτύων, όπως για παράδειγμα το γεγονός ότι μερικά από τα νευρωνικά δίκτυα έχουν πολλά τοπικά ελάχιστα. Επίσης, η εύρεση του αριθμού των νευρώνων που θα χρειαστούν

για μια εργασία είναι ένα άλλο θέμα που καθορίζει αν έχει επιτευχθεί το βέλτιστο νευρωνικό δίκτυο. Ένα σημαντικό ακόμα σημείο που πρέπει να αναφέρουμε εδώ είναι ότι ακόμα και αν οι λύσεις των νευρωνικών δικτύων τείνουν να συγκλίνουν αυτό μπορεί να μην έχει σαν αποτέλεσμα μια και μοναδική λύση (Mitchell, 1997).
Ακολουθεί ένα παράδειγμα, στο οποίο το δεδομένα παρουσιάζονται στο παρακάτω διάγραμμα και γίνεται προσπάθεια ταξινόμησής τους και στο οποίο υπάρχουν πολλά υπερεπίπεδα που μπορούν να χρησιμοποιηθούν για να ταξινομηθούν τα δεδομένα. Το πρόβλημα έγκειται στο να βρεθεί το καλύτερο από αυτά.



Εικόνα 3. - τα υπερεπίπεδα που μπορούν να ταξινομήσουν τα δεδομένα.
Εδώ προκύπτει η ανάγκη για χρήση SVM τεχνικής.
Από την Εικόνα 3, φαίνεται ότι υπάρχουν πολλοί γραμμικοί ταξινομητές (υπερεπίπεδα) που διαχωρίζουν τα δεδομένα. Όμως, μόνο ένας από αυτούς μπορεί να επιτύχει τον μέγιστο διαχωρισμό. Ο λόγος που χρειαζόμαστε τον μέγιστο δυνατό διαχωρισμό, είναι επειδή αν χρησιμοποιήσουμε ένα υπερεπίπεδο για την ταξινόμηση των δεδομένων, μπορεί να καταλήξει πλησιέστερα σε ένα μόνο σύνολο συνόλων δεδομένων σε σύγκριση με άλλα και αυτό δεν είναι επιθυμητό. Επομένως μια λύση είναι η ιδέα του ταξινομητή μέγιστου περιθωρίου. Η επόμενη απεικόνιση δίνει παράδειγμα με τον ταξινομητή μέγιστου περιθωρίου που παρέχει μια λύση στο πρόβλημα που αναφέρθηκε παραπάνω (Cristianini and Shawe-Taylor, 2000).



Εικόνα 4. - Αναπαράσταση των γραμμικών μηχανών υποστήριξης διανυσμάτων.

Το μέγιστο περιθώριο δίνεται από την ακόλουθη σχέση (Burges B., 1998; Cristianini and Shawe-Taylor, 2000)

$$margin \equiv arg \min_{x \in D} d(x) = arg \min_{x \in D} \frac{|\boldsymbol{x} \cdot \boldsymbol{w} + b|}{\sqrt{\sum_{i=1}^{d} w_i^2}}$$

Η Εικόνα 4 απεικονίζει τον μέγιστο γραμμικό ταξινομητή με το μέγιστο εύρος. Μία άλλη ενδιαφέρουσα ερώτηση στην οποία πρέπει να απαντήσουμε είναι γιατί ζητείται το μέγιστο περιθώριο. Υπάρχουν αρκετές καλές επεξηγήσεις σε αυτό το ερώτημα, που περιλαμβάνουν την καλύτερη εμπειρική απόδοση. Ένας άλλος καλός λόγος είναι ότι ακόμα και αν έχει συμβεί σφάλμα στην τοποθεσία του συνόρου-περιθωρίου, αυτό δίνει ελάχιστη πιθανότητα πρόκλησης λανθασμένης ταξινόμησης. Ένα άλλο πλεονέκτημα θα ήταν η αποφυγή του τοπικού ελαχίστου και η καλύτερη ταξινόμηση.

Στο σημείο αυτό θα παρουσιάσουμε με μαθηματικές σχέσεις πως λειτουργεί μία Μηχανή Υποστήριξης Διανυσμάτων και συγκεκριμένα την πιο κλασική περίπτωση, δηλαδή μια γραμμική Μηχανή Υποστήριξης Διανυσμάτων. Οι στόχοι των SVM είναι να διαχωρίσουν τα δεδομένα με την χρήση ενός υπερεπίπεδου και να το επεκτείνουν σε μη γραμμικά σύνορα χρησιμοποιώντας κάποιο τέχνασμα του πυρήνα (kernel trick) (Cristianini and Shawe-Taylor, 2000; Mitchell, 1997). Για τον υπολογισμό της Μηχανής Διανύσματος Υποστήριξης βλέπουμε ότι ο στόχος είναι η σωστή ταξινόμηση των δεδομένων. Για τους μαθηματικούς υπολογισμούς έχουμε:

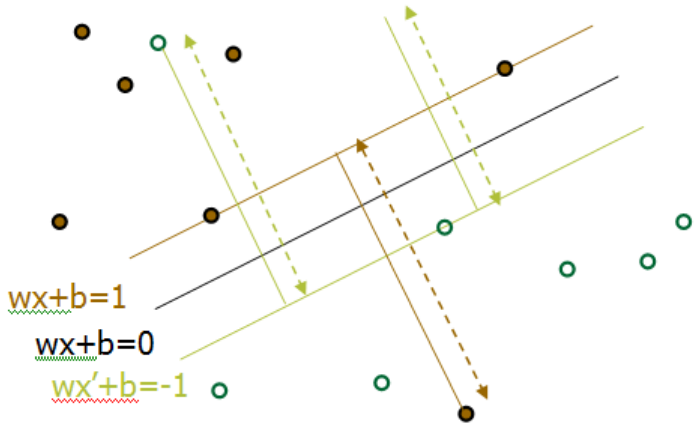
[a] Εάν $Y_i = +1;$ $\quad wx_i + b \geq 1$

[b] Εάν $Y_i = -1;$ $wx_i + b \leq 1$

[c] Για κάθε $i;$ $y_i (w_i + b) \geq 1$

όπου *x* είναι ένα διάνυσμα σημείο και *w* είναι το βάρος και είναι επίσης διάνυσμα.

Ανάμεσα σε όλα τα πιθανά υπερεπίπεδα, η SVM τεχνική επιλέγει αυτό με την μεγαλύτερη δυνατή απόσταση, αν τα δεδομένα εκπαίδευσης είναι καλά και κάθε διάνυσμα ελέγχου βρίσκεται σε ακτίνα r από το διάνυσμα εκπαίδευσης. Το επιλεγμένο υπερεπίπεδο πρέπει να βρίσκεται όσο πιο απόμακρα γίνεται από τα δεδομένα (Lewis, 2004). Αυτό το επιθυμητό υπερεπίπεδο που μεγιστοποιεί το περιθώριο, διχοτομεί τις γραμμές μεταξύ των πλησιέστερων σημείων στην κυρτή θήκη των δύο συνόλων δεδομένων. Έτσι έχουμε τα [a], [b] και [c] όπως απεικονίζονται παρακάτω.



Εικόνα 5. - Στην παραπάνω εικόνα παρουσιάζονται τα διάφορα υπερεπίπεδα.

Η απόσταση του πλησιέστερου σημείου πάνω στο υπερεπίπεδο στην αρχή μπορεί να βρεθεί μεγιστοποιώντας το *x*, καθώς το *x* βρίσκεται πάνω στο υπερεπίπεδο. Για τα σημεία που βρίσκονται στην άλλη μεριά ισχύει ένα παρόμοιο σενάριο. Έτσι, λύνοντας και αφαιρώντας τις δύο διαστάσεις παίρνουμε την αθροισμένη απόσταση από το διαχωριστικό υπερεπίπεδο στα πλησιέστερα σημεία. Το μέγιστο περιθώριο είναι το εξής:

$$\text{MaximumMargin}=M=2/||w||$$

Η μεγιστοποίηση του περιθωρίου είναι ίδια με την ελαχιστοποίηση (Cristianini and Shawe-Taylor, 2000). Θεωρούμε ότι έχουμε ένα τετραγωνικό πρόβλημα βελτιστοποίησης που χρειάζεται να λυθεί ως προς *w* και ως προς b. Για να επιτευχθεί αυτό πρέπει να βελτιστοποιηθεί η τετραγωνική συνάρτηση με γραμμικούς περιορισμούς. Η λύση περιλαμβάνει τη κατασκευή ενός δυικού προβλήματος όπου υπάρχει ένας Langlier's πολλαπλασιαστής α_i. Χρειάζεται να υπολογιστούν οι τιμές των *w* και b για τις οποίες ελαχιστοποιείται η παρακάτω παράσταση

$$\Phi(w) = \frac{1}{2}|w'||w| \text{ για όλα τα } \{(x_i, y_i)\}: y_i(w * x_i + b) \geq 1.$$

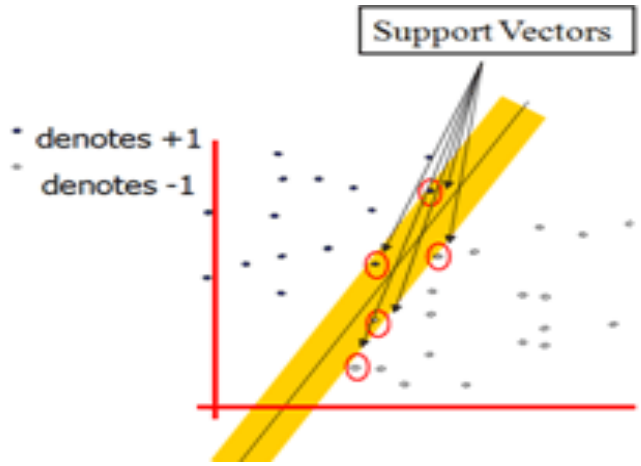
Μετά την επίλυση του προβλήματος προκύπτει ότι:

$$w = \sum a_i * x_i, b = y_k - w * x_k$$

για όλα τα x_k για τα οποία ισχύει $\alpha_k \neq 0$.
Η συνάρτηση ταξινόμησης είναι της μορφής:

$$f(x) = \sum a_i y_i x_i * x + b.$$

Στην εικόνα που ακολουθεί παρουσιάζονται τα διανύσματα υποστήριξης.



Εικόνα 6.: Αναπαράσταση των διανυσμάτων υποστήριξης

3.4 Αναπαράσταση των Μηχανών Διανυσμάτων Υποστήριξης

Σε αυτή τη παράγραφο γίνεται παρουσίαση της QP διατύπωσης (Burges B., 1998; Cristianini and Shawe-Taylor, 2000; Lewis, 2004; Vapnik, 1998) για την SVM ταξινόμηση.

Για την SV ταξινόμηση ισχύει:

$$y_i f(x_i) \geq 1 - \xi_i$$

και

$$\min_{f,\xi_i} \|f\|_K^2 + C \sum_{i=1}^{l} \xi_i$$

για όλα τα i για τα οποία $\xi_i \geq 0$.

Για τη SVM ταξινόμηση και δυική φόρμουλα ισχύει:

$$\min_{a_i} \sum_{i=1}^{l} a_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} a_i a_j y_i y_j K(x_i, x_j)$$

και

$$0 \leq a_i \leq C \ \forall i \text{ ώστε } \sum_{i=1}^{l} a_i y_i = 0$$

Οι μεταβλητές ξ_i καλούνται χαλαρές μεταβλητές και μετρούν το σφάλμα στο σημείο (x_i, y_i). Η εκπαίδευση μιας Μηχανής Υποστήριξης Διανυσμάτων γίνεται πρόκληση όταν ο αριθμός των σημείων εκπαίδευσης είναι μεγάλος και μέχρι πρότινος έχει προταθεί ένας αριθμός μεθόδων για γρήγορη εκπαίδευση των Μηχανών Υποστήριξης Διανυσμάτων (Burges B., 1998; Cristianini and Shawe-Taylor, 2004; Vapnik, 1998).

3.5 Soft Margin Ταξινομητής

Σε ένα ρεαλιστικό πρόβλημα δεν είναι πιθανό να διαχωριστούν τα δεδομένα με μια ξεχωριστή γραμμή μέσα στον χώρο. Μπορεί να υπάρχει ένα κυρτό σύνορο απόφασης ή ακόμα μπορεί να υπάρχει ένα υπερεπίπεδο που διαχωρίζει τα δεδομένα, αλλά αυτό ίσως δεν είναι επιθυμητό αν τα δεδομένα έχουν 'θόρυβο' (noise) μέσα σε αυτό. Είναι καλύτερο λοιπόν για το ομαλό περιθώριο να αγνοήσει μερικά σημεία δεδομένων από το να είναι κυρτό ή να μπει σε βρόχο, γύρω από τις ακραίες τιμές. Εδώ λοιπόν εισάγονται οι χαλαρές μεταβλητές για να διαχειριστούν μία τέτοια διαφορετική κατάσταση (Burges B., 1998; Lewis, 2004). Η νέα σχέση που παρουσιάζει πλέον το πρόβλημα είναι η παρακάτω:

$$y_i(w^{'} x + b) \geq 1 - S_k$$

Αυτό επιτρέπει σε ένα σημείο να έχει μικρή απόσταση, S_k, στην λανθασμένη πλευρά του υπερεπιπέδου χωρίς να παραβιάζεται ο περιορισμός. Σε αυτή τη περίπτωση μπορεί να προκύπτουν τεράστιες χαλαρές μεταβλητές που επιτρέπουν σε οποιαδήποτε γραμμή να διαχωρίζει τα δεδομένα. Σε ένα τέτοιο σενάριο εισάγεται η Lagrangian μεταβλητή η οποία ποινικοποιεί τις μεγάλες χαλαρές μεταβλητές με ένα τρόπο ο οποίος περιγράφεται από τη παρακάτω σχέση:

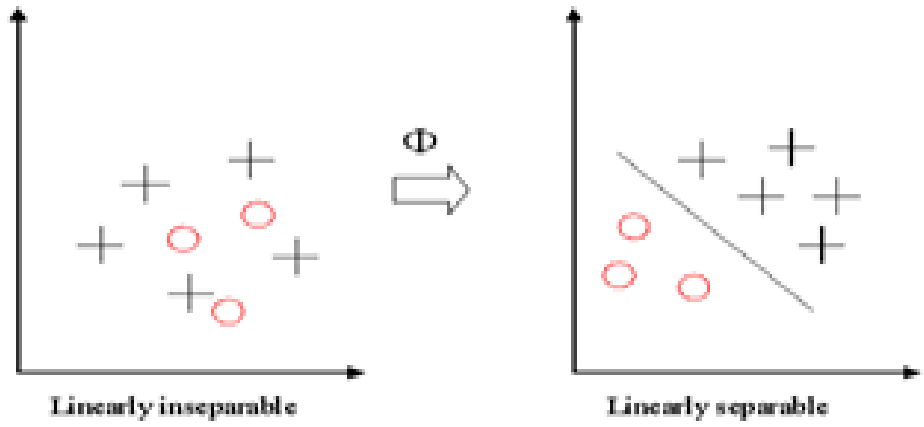
$$minL = \frac{1}{2} w^{'} w - \sum \lambda_k (y_k (w^{'} x_k + b) + S_k - 1) + a \sum S_k$$

Μειώνοντας τώρα το a επιτρέπεται πιο πολλά δεδομένα να βρίσκονται στην λανθασμένη πλευρά του υπερεπιπέδου και θα αντιμετωπιστούν ως ακραίες τιμές που δίνουν πιο χαλαρό σύνορο απόφασης (Lewis, 2004).

3.6 Τέχνασμα του πυρήνα

- Ορισμός του πυρήνα (kernel):

Όταν τα δεδομένα είναι γραμμικά, μπορεί να χρησιμοποιηθεί ένα διαχωριστικό υπερεπίπεδο για να χωρίσει τα δεδομένα. Παρόλα αυτά είναι συχνή η περίπτωση που τα δεδομένα απέχουν πολύ από την γραμμικότητα και τα σύνολα δεδομένων είναι μη διαχωρίσιμα. Σε μια τέτοια περίπτωση χρησιμοποιούνται οι πυρήνες για να χαρτογραφήσουν τα μη γραμμικά δεδομένα εισόδου σε ένα χώρο υψηλής διάστασης. Η νέα χαρτογράφηση καταλήγει να είναι γραμμικά διαχωρίσιμη. Στην εικόνα που ακολουθεί φαίνεται η απεικόνιση μια τέτοιας χαρτογράφησης.



Εικόνα 7. Απεικόνιση χαρτογράφησης η οποία καταλήγει να είναι γραμμικά διαχωρίσιμη.

Αυτή η χαρτογράφηση καθορίζεται από τον πυρήνα:

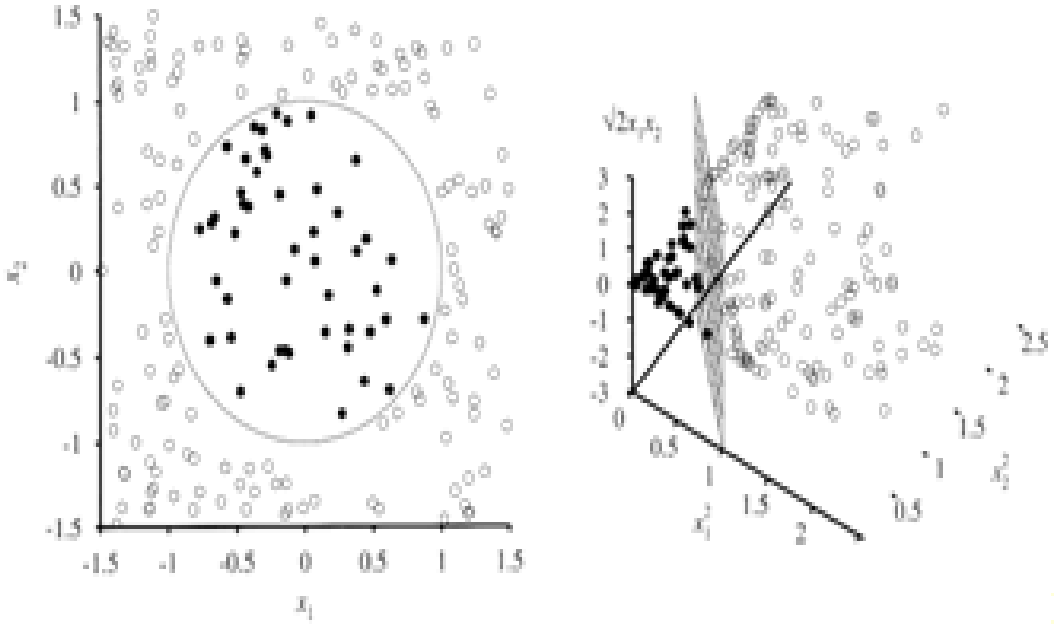
$$K(x, y) = \Phi(x) \cdot \Phi(y).$$

- Ορισμός του χώρου χαρακτηριστικών (feature space):

Ο μετασχηματισμός των δεδομένων μέσα στο χώρο των χαρακτηριστικών καθιστά εύκολο το να καθοριστεί ένα μέτρο ομοιότητας με βάση το βαθμωτό γινόμενο ως εξής:

$$\langle x_1 \cdot x_2 \rangle \leftarrow K(x_1, x_2) = \langle \Phi(x_1) \cdot \Phi(x_2) \rangle$$

Εφόσον ο χώρος των χαρακτηριστικών έχει επιλεγεί κατάλληλα, η αναγνώριση του προτύπου θα είναι εύκολη.



Εικόνα 8. Αναπαράσταση του χώρου χαρακτηριστικών.

Όσον αφορά το τέχνασμα του πυρήνα, παρατηρείται ότι όταν τα *w, b* είναι γνωστά το πρόβλημα λύνεται για ένα απλό γραμμικό σενάριο στο οποίο τα δεδομένα διαχωρίζονται με ένα υπερεπίπεδο. Το τέχνασμα του πυρήνα επιτρέπει στην SVM τεχνική να σχηματίσει μη γραμμικά σύνορα. Τα βήματα του τεχνάσματος πυρήνα είναι τα παρακάτω (Lewis, 2004; Burges B., 1998):

a. Ο αλγόριθμος εκφράζεται με την χρήση των εσωτερικών γινομένων των συνόλων δεδομένων. Αυτό καλείται δυϊκό πρόβλημα.
b. Τα αρχικά δεδομένα διέρχονται από μη γραμμικούς χάρτες ώστε να δημιουργήσουν νέα δεδομένα σε σχέση με τις νέες διαστάσεις, προσθέτοντας ένα γινόμενο ανά ζεύγη από μερικά από τα αρχικά δεδομένα σε κάθε διάνυσμα δεδομένων.
c. Έπειτα αντί να χρησιμοποιηθεί ένα εσωτερικό γινόμενο σε αυτά τα νέα δεδομένα, χρησιμοποιούνται μεγαλύτερα διανύσματα αποθηκευμένα σε πίνακες και αργότερα πραγματοποιείται μία διαδικασία αναζήτησης μέσα στον πίνακα. Τέλος, παρουσιάζεται ένα βαθμωτό γινόμενο των δεδομένων μετά την μη γραμμική χαρτογράφηση τους. Αυτή η συνάρτηση είναι η συνάρτηση πυρήνα. Παρακάτω δίνονται περισσότερες πληροφορίες σχετικά με τις συναρτήσεις πυρήνα.

- Τέχνασμα του πυρήνα: Δυϊκό πρόβλημα

Πρώτα, γίνεται η μετατροπή του προβλήματος βελτιστοποίησης στην δυϊκή μορφή, στην οποία προσπαθούμε να ελαχιστοποιήσουμε το *w*, και η Lagrangian τώρα είναι μια συνάρτηση λ_i. Υπάρχει μαθηματική λύση για αυτό το πρόβλημα, στην οποία όμως δεν θα αναφερθούμε διεξοδικά στη παρούσα εργασία. Συνοπτικά, για να επιλυθεί το πρόβλημα πρέπει να μεγιστοποιηθεί το L_D σε σχέση με το λ_i. Το Δυϊκό πρόβλημα απλοποιεί την βελτιστοποίηση και επιτυγχάνει σαν αποτέλεσμα την

απόκτηση του βαθμωτού γινόμενου από αυτό (Burges C., 1998; Cristianini and Shawe-Taylor, 2000; Lewis, 2004).

- Τέχνασμα του πυρήνα: Περιληπτική περιγραφή των εσωτερικών γινομένων

Το βαθμωτό γινόμενο των μη γραμμικά χαρτογραφημένων δεδομένων μπορεί να είναι δαπανηρό. Το τέχνασμα του πυρήνα επιλέγει μια κατάλληλη συνάρτηση που αντιστοιχεί στο βαθμωτό γινόμενο κάποιας μη γραμμικής χαρτογράφησης (Burges C., 1998; Cristianini and Shawe-Taylor, 2000; Lewis, 2004). Μερικές από τις πιο συχνά επιλεγμένες συναρτήσεις πυρήνα περιγράφονται παρακάτω. Κάθε φορά επιλέγεται εμπειρικά ένας συγκεκριμένος πυρήνας στο σύνολο ελέγχου, επιλέγοντας τον σωστό πυρήνα που θα ενίσχυε την απόδοση της SVM τεχνικής, με βάση το πρόβλημα ή την εφαρμογή που αντιμετωπίζει ο πειραματιστής.

3.7 Συναρτήσεις πυρήνα

Η βασική ιδέα στην οποία στηρίζεται μια συνάρτηση πυρήνα είναι να επιτρέψει την εκτέλεση των λειτουργιών στο χώρο εισόδου αντί στον υψηλής διάστασης χώρο χαρακτηριστικών. Έτσι, το εσωτερικό γινόμενο δεν χρειάζεται να αξιολογηθεί στον χώρο χαρακτηριστικών. Η συνάρτηση πυρήνα πρέπει να εκτελεί την χαρτογράφηση των γνωρισμάτων του χώρου εισόδου στον χαρακτηριστικό χώρο. Η συνάρτηση πυρήνα παίζει σημαντικό ρόλο στην SVM τεχνική και στην απόδοσή της και βασίζεται στην αναπαραγωγή των πυρήνων σε χώρους Hilbert (Lewis, 2004; Aizerman et al.,1964; Aronszajn, 1950; Heckman, 1997):

$$K(x, x') = \langle \varphi(x), \varphi(x''''') \rangle,$$

Αν η K είναι μια συμμετρική θετικά ορισμένη συνάρτηση, που ικανοποιεί τις συνθήκες του Mercer, οι οποίες είναι,

$$K(x, x') = \sum_{m}^{\infty} a_m \varphi_m(x) \varphi_m(x'), \quad a_m \geq 0,$$

$$\int \int K(x, x') g(x) g(x') dx dx' > 0, \qquad g \in L_2$$

τότε ο πυρήνας αντιπροσωπεύει ένα επιτρεπτό εσωτερικό γινόμενο στον χώρο χαρακτηριστικών. Το σύνολο εκπαίδευσης είναι γραμμικά διαχωρίσιμο στον χώρο χαρακτηριστικών. Αυτό καλείται και ως «τέχνασμα του πυρήνα» (Cristianini and Shawe-Taylor, 2000; Lewis, 2004).

Οι διαφορετικές συναρτήσεις πυρήνα παρατίθενται παρακάτω (Cristianini and Shawe-Taylor, 2000; Lewis, 2004):

i. Πολυωνυμική: Η πολυωνυμική χαρτογράφηση είναι μια γνωστή μέθοδος για μη γραμμική μοντελοποίηση. Ο δεύτερος πυρήνας είναι προτιμότερος, καθώς αποφεύγει προβλήματα με τον Hessian να γίνεται μηδέν.

$$K(x, x') = \langle x, x' \rangle^d$$

$$K(x, x') = (\langle x, x' \rangle + 1)^d$$

ii. Gaussian Radial Basis Function: Οι radial basis συναρτήσεις έχουν συνήθως Gaussian μορφή

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$$

iii. Exponential Radial Basis Function: Μια radial basis συνάρτηση παράγει μια κατά τμήματα γραμμική λύση που μπορεί να είναι ελκυστική όταν είναι αποδεκτές οι ασυνέχειες

$$K(x, x'') = \exp\left(-\frac{\|x - x'\|}{2\sigma^2}\right)$$

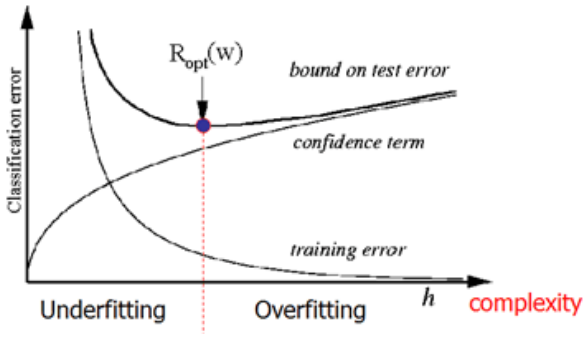
iv. Multi-layer Perceptron: Ο MLP με ένα μόνο κρυμμένο επίπεδο έχει μια έγκυρη αναπαράσταση πυρήνα

$$K(x, x') = \tanh(p(x, x') + e)$$

Υπάρχουν πολλές ακόμα συναρτήσεις πυρήνα που περιλαμβάνουν Fourier, splines, B-splines, πρόσθετους πυρήνες και γινόμενα τανυστών (Cristianini and Shawe-Taylor, 2000).

3.8 Έλεγχος πολυπλοκότητας στην SVM τεχνική: Εξισορρόπηση παραγόντων

Η SVM τεχνική έχει την δυνατότητα να προσεγγίσει οποιαδήποτε δεδομένα εκπαίδευσης και γενικεύει καλύτερα σε δοσμένα σύνολα δεδομένων. Η πολυπλοκότητα όσον αφορά τις συναρτήσεις πυρήνα επηρεάζει την απόδοση των νέων συνόλων δεδομένων (Cristianini and Shawe-Taylor, 2000). Η SVM τεχνική υποστηρίζει παραμέτρους για τον έλεγχο της πολυπλοκότητας και πάνω από όλα η SVM τεχνική δεν δίνει πληροφορίες για το πώς ρυθμίζονται οι παράμετροι αυτοί. Για το λόγο αυτό ο πειραματιστής θα πρέπει να είναι ικανός να καθορίσει τις τιμές των παραμέτρων αυτών χρησιμοποιώντας την τεχνική της διασταυρωμένης επικύρωσης (cross-validation) στα δοσμένα σύνολα δεδομένων (Mitchell, 1997).



Εικόνα 9. Το παραπάνω διάγραμμα παρουσιάζει μια απεικόνιση για το τρόπο με τον οποίο ελέγχεται η πολυπλοκότητα.

3.9 Η χρήση της SVM τεχνικής για τη ταξινόμηση

Η SVM τεχνική είναι μια χρήσιμη τεχνική για τη ταξινόμηση δεδομένων. Αν και θεωρείται ότι τα Νευρωνικά Δίκτυα είναι πιο εύκολα στην χρήση από ότι τα SVMs, μερικές φορές φέρνουν μη ικανοποιητικά αποτελέσματα. Σε μια διαδικασία ταξινόμησης συμμετέχουν συνήθως δεδομένα εκπαίδευσης και δεδομένα δοκιμής τα οποία αποτελούνται από διάφορα παραδείγματα δεδομένων (Duda and Hart, 1973). Κάθε παράδειγμα δεδομένων στο σύνολο εκπαίδευσης περιέχει μία τιμή στόχου και διάφορα γνωρίσματα-μεταβλητές. Ο στόχος της SVM τεχνικής είναι να παράγει ένα μοντέλο που να προβλέπει την τιμή στόχο των διάφορων δεδομένων στο σύνολο δοκιμής, δοθέντων μόνο των γνωρισμάτων-μεταβλητών (Cristianini and Shawe-Taylor, 2000).

Η ταξινόμηση στη SVM τεχνική είναι ένα παράδειγμα εκμάθησης με επίβλεψη. Οι γνωστές ετικέτες υποδεικνύουν αν το σύστημα λειτουργεί με σωστό τρόπο ή όχι. Αυτή η πληροφορία επηρεάζει σημαντικά την επιθυμητή απόκριση, επικυρώνοντας την ακρίβεια του συστήματος, ή χρησιμοποιείται για να βοηθήσει το σύστημα να μάθει να δρα σωστά. Ένα βήμα στην SVM ταξινόμηση περιλαμβάνει την αναγνώριση των χαρακτηριστικών, μια διαδικασία στενά συνδεδεμένη με τις γνωστές κλάσεις. Αυτό καλείται επιλογή χαρακτηριστικών ή εξαγωγή χαρακτηριστικών. Η επιλογή χαρακτηριστικών και η SVM ταξινόμηση μαζί μπορούν να χρησιμοποιηθούν ακόμα και όταν η πρόβλεψη αγνώστων δειγμάτων δεν είναι απαραίτητη. Μπορούν να χρησιμοποιηθούν για την αναγνώριση βασικών συνόλων που περιλαμβάνονται σε οποιαδήποτε διαδικασία διαχωρισμού κλάσεων (Cristianini and Shawe-Taylor, 2000).
.

3.10 Η χρήση της SVM τεχνικής για την παλινδρόμηση

Οι Μηχανές Υποστήριξης Διανυσμάτων μπορούν να εφαρμοστούν σε προβλήματα παλινδρόμησης με την εισαγωγή μιας εναλλακτικής συνάρτησης απώλειας

(Cristianini and Shawe-Taylor, 2000; Smola, 1996). Η συνάρτηση απώλειας πρέπει να τροποποιηθεί ώστε να περιλαμβάνει ένα μέτρο απόστασης. Η παλινδρόμηση μπορεί να είναι γραμμική ή μη γραμμική. Τα γραμμικά μοντέλα αποτελούνται κυρίως από τις ακόλουθες συναρτήσεις απώλειας, e-intensive συναρτήσεις απώλειας, τετραγωνικές και Huber συναρτήσεις απώλειας.

Παρόμοια με τα προβλήματα ταξινόμησης, ένα μη γραμμικό μοντέλο συνήθως απαιτείται ώστε να μοντελοποιήσει επαρκώς τα δεδομένα. Μία μη γραμμική χαρτογράφηση μπορεί να χρησιμοποιηθεί ώστε να χαρτογραφηθούν τα δεδομένα σε ένα υψηλής διάστασης χώρο χαρακτηριστικών, όπου εκτελείται γραμμική παλινδρόμηση. Η προσέγγιση του πυρήνα χρησιμοποιείται ξανά για να κατευθύνει την διάσταση. Στην μέθοδο παλινδρόμησης υπάρχουν θεωρίες που βασίζονται σε προηγούμενη γνώση του προβλήματος (prior knowledge) και την κατανομή του θορύβου. Σε περιπτώσεις απουσίας αυτών των πληροφοριών, η εύρωστη συνάρτηση απώλειας του Huber έχει αποδειχτεί μια καλή εναλλακτική (Cristianini and Shawe-Taylor, 2000; Cortes and Vapnik, 1995).

3.11 Εφαρμογές των Μηχανών Υποστήριξης Διανυσμάτων

Οι Μηχανές Υποστήριξης Διανυσμάτων έχουν αποδειχτεί επιτυχημένες όταν έχουν χρησιμοποιηθεί σε προβλήματα ταξινόμησης προτύπων. Η εφαρμογή της προσέγγισης των Μηχανών Υποστήριξης Διανυσμάτων σε ένα πρακτικό πρόβλημα περιλαμβάνει επίλυση ενός αριθμού ερωτήσεων, οι οποίες προκύπτουν από το καθορισμό του προβλήματος καθώς και στον σχεδιασμό του. Μία από τις μεγαλύτερες προκλήσεις είναι αυτή της επιλογής ενός κατάλληλου πυρήνα για την δοθείσα εφαρμογή (Burges C., 1998). Υπάρχουν κλασικά σταθερές επιλογές όπως ο Gaussian ή ο πολυωνυμικός πυρήνας, οι οποίες αποτελούν τις εξ' ορισμού επιλογές, αλλά αν αυτές αποδειχθούν αναποτελεσματικές ή αν τα δεδομένα εισόδου είναι διακριτές δομές, τότε θα χρειαστούν πιο περίπλοκοι πυρήνες. Με τον εν δυνάμει καθορισμό ενός χώρου χαρακτηριστικών, ο πυρήνας παρέχει την γλώσσα περιγραφής που χρησιμοποιείται από την μηχανή για την προβολή των δεδομένων. Από την στιγμή που η επιλογή πυρήνα και το κριτήριο βελτιστοποίησης έχουν καθοριστεί, τα σημαντικά συστατικά του συστήματος είναι πλέον γνωστά (Cristianini and Shawe-Taylor, 2000).

Παρακάτω υπάρχουν μερικά παραδείγματα. Η εργασία της κατηγοριοποίησης κειμένου είναι η ταξινόμηση των φυσικών εγγράφων κειμένου σε ένα σταθερό αριθμό προκαθορισμένων κατηγοριών που βασίζονται στο περιεχόμενό τους. Εφόσον ένα έγγραφο μπορεί να εκχωρηθεί σε περισσότερες από μία κατηγορίες, αυτό μπορεί να μην αντιμετωπιστεί ως ένα πρόβλημα ταξινόμησης πολλών κλάσεων, αλλά μπορεί να θεωρηθεί ως σειρά δυαδικών προβλημάτων ταξινόμησης, ένα για κάθε κατηγορία. Μία από τις κλασικές αναπαραστάσεις κειμένου, για τους σκοπούς της ανάκτησης πληροφοριών, παρέχει μία ιδανική χαρτογράφηση χαρακτηριστικών για την

κατασκευή ενός πυρήνα Mercer (Osuna et al., 1997). Πράγματι, οι πυρήνες με κάποιο τρόπο ενσωματώνουν ένα μέτρο ομοιότητας μεταξύ των διάφορων περιπτώσεων, γεγονός που καθιστά λογικό το γεγονός ότι ήδη οι ειδικοί που εργάζονται στο συγκεκριμένο πεδίο εφαρμογής έχουν προσδιορίσει έγκυρα μέτρα ομοιότητας, ειδικά σε περιοχές όπως η ανάκτηση πληροφοριών και η παραγωγή μοντέλων.

Κλασικές προσεγγίσεις ταξινόμησης αποδίδουν φτωχά κατά τις άμεσες διεργασίες τους, εξαιτίας της υψηλής διάστασης των δεδομένων, αλλά οι Μηχανές Διανυσμάτων Υποστήριξης μπορούν να αποφύγουν τις παγίδες των υψηλής διάστασης αναπαραστάσεων (Lewis, 2004). Μια παρόμοια προσέγγιση με τις τεχνικές που περιγράφηκαν για την κατηγοριοποίηση κειμένου μπορεί να χρησιμοποιηθεί για την ταξινόμηση εικόνων. Η πρώτη εργασία στον πραγματικό κόσμο στην οποία ελέγχθηκαν οι Μηχανές Υποστήριξης Διανυσμάτων ήταν το πρόβλημα της αναγνώρισης του γραφικού χαρακτήρα, μάλιστα και σε περιπτώσεις δεδομένων πολλών κλάσεων. Ένα ενδιαφέρον θέμα αποτελεί η σύγκριση των Μηχανών Υποστήριξης Διανυσμάτων με άλλους ταξινομητές αλλά και η σύγκριση μεταξύ των Μηχανών Υποστήριξης Διανυσμάτων (Stitson and Weston, 1996). Από μελέτες που έχουν πραγματοποιηθεί μέχρι τώρα αποδεικνύεται ότι έχουν περίπου την ίδια απόδοση και επιπλέον μοιράζονται τα περισσότερα διανύσματα στήριξης, ανεξάρτητα από τον επιλεγμένο πυρήνα. Είναι αξιοσημείωτο το γεγονός ότι οι Μηχανές Υποστήριξης Διανυσμάτων μπορούν να αποδώσουν τόσο καλά όσο και συστήματα τα οποία περιλαμβάνουν λεπτομερή εκ των προτέρων γνώση (Osuna E. et al., 1997).

3.12 Πλεονεκτήματα και μειονεκτήματα των Μηχανών Υποστήριξης Διανυσμάτων

Τα κύρια πλεονεκτήματα των Μηχανών Διανυσμάτων Υποστήριξης είναι ότι η εκπαίδευση είναι σχετικά εύκολη και ότι δεν έχουν τοπικά μέγιστα όπως τα νευρωνικά δίκτυα. Λειτουργεί σχετικά καλά σε υψηλής διάστασης δεδομένα και η εξισορρόπηση μεταξύ της πολυπλοκότητας των ταξινομητών και του σφάλματος μπορεί να ελεγχθεί. Το βασικό ελάττωμα των Μηχανών Υποστήριξης Διανυσμάτων περιλαμβάνει την ανάγκη για μια καλή συνάρτηση πυρήνα.

3.13 Συμπέρασμα

Συμπερασματικά, οι Μηχανές υποστήριξης διανυσμάτων βασίζονται στη στατιστική θεωρία εκμάθησης. Μπορούν να χρησιμοποιηθούν για την πρόβλεψη μελλοντικών δεδομένων μέσω διαδικασιών εκμάθησης (Osuna et al, 1997). Οι Μηχανές υποστήριξης διανυσμάτων εκπαιδεύονται επιλύοντας τετραγωνικά προβλήματα βελτιστοποίησης με περιορισμούς. Εκτελούν τη χαρτογράφηση των δεδομένων εισόδου σε υψηλής διάστασης χώρους με την χρήση μη γραμμικών συναρτήσεων

βάσης. Υπάρχει η δυνατότητα να χρησιμοποιηθούν για την εκπαίδευση μιας ποικιλίας αναπαραστάσεων όπως είναι τα νευρωνικά δίκτυα, οι splines και οι πολυωνυμικοί εκτιμητές και μάλιστα υπάρχει μία μοναδική βέλτιστη λύση για τη κάθε επιλογή των SVM παραμέτρων (Burges B., 1998). Σε αυτό το σημείο διαφέρουν σημαντικά άλλες μηχανές εκμάθησης, όπως είναι τα πρότυπα νευρωνικά δίκτυα που εκπαιδεύονται με τροφοδότηση προς τα πίσω-διάδοσης. Εν συντομία, η ανάπτυξη των Μηχανών υποστήριξης διανυσμάτων προσφέρει μια νέα αντίληψη στις τεχνικές εκμάθησης. Τα τέσσερα κυριότερα χαρακτηριστικά των SVMs είναι η δυαδικότητα, οι πυρήνες, η κυρτότητα και η σπανιότητα (Burges B., 1998).

Οι Μηχανές υποστήριξης διανυσμάτων αποτελούν μία από τις καλύτερες προσεγγίσεις με στόχο τη μοντελοποίηση των δεδομένων. Συνδυάζουν τον έλεγχο γενίκευσης σαν μια τεχνική ελέγχου της διάστασης. Η χαρτογράφηση του πυρήνα αποτελεί μια κοινή βάση για τις περισσότερες από τις συνήθεις αρχιτεκτονικές μοντέλου, επιτρέποντας συγκρίσεις (Cristianini and Shawe-Taylor, 2000). Στα προβλήματα ταξινόμησης ο έλεγχος γενίκευσης γίνεται με την μεγιστοποίηση του περιθωρίου, που αντιστοιχεί σε ελαχιστοποίηση του διανύσματος βάρους σε ένα κανονικό πλαίσιο. Η λύση βρίσκεται σαν ένα σύνολο διανυσμάτων στήριξης που μπορεί να είναι αραιό. Η ελαχιστοποίηση του διανύσματος βάρους χρησιμοποιείται σαν κριτήριο στα προβλήματα παλινδρόμησης, με μια τροποποιημένη συνάρτηση απώλειας. Μελλοντικές Κατευθύνσεις για θέματα που πρέπει να ερευνηθούν σχετικά με τις Μηχανές Υποστήριξης Διανυσμάτων περιλαμβάνουν:

i. Την εύρεση τεχνικών επιλογής της συνάρτησης πυρήνα και πρόσθετης ικανότητα ελέγχου.
ii. Ανάπτυξη πυρήνων με σταθερότητα.
iii. Νέες τεχνικές εκμάθησης των SVMs, οι οποίες προτάθηκαν πρόσφατα από τον Vapnik (Vapnik, 2006).

3.14 Σύνδεση των Μηχανών Διανυσμάτων Υποστήριξης με τα Νευρωνικά Δίκτυα

Τις Μηχανές Διανυσμάτων Υποστήριξης ή αλλιώς Support Vector Machines (SVMs) μπορούμε να τις δούμε ως ένα καινούργιο τρόπο εκπαίδευσης των νευρωνικών δικτύων τροφοδότησης προς τα εμπρός διάδοσης (feed-forward). Πιο συγκεκριμένα, μπορούμε να χρησιμοποιήσουμε τον αλγόριθμο εκμάθησης των διανυσμάτων υποστήριξης για να δημιουργήσουμε τους ακόλουθους τρεις τύπους μηχανών εκμάθησης μεταξύ άλλων:

i. Polynomial learning machines,
ii. Radial-basis function networks,
iii. Two-layer perceptrons (δηλαδή με ένα hidden layer).

Δηλαδή, για κάθε ένα από αυτά τα δίκτυα τροφοδότησης προς τα εμπρός διάδοσης μπορούμε να χρησιμοποιήσουμε τον αλγόριθμο εκμάθησης των διανυσμάτων υποστήριξης για να υλοποιήσουμε μια διαδικασία εκμάθησης χρησιμοποιώντας ένα σύνολο από δεδομένα, αποφασίζοντας αυτόματα τον απαιτούμενο αριθμό των κρυμμένων μονάδων. Με άλλα λόγια ενώ ο αλγόριθμος προς τα πίσω-διάδοσης είναι σχεδιασμένος ειδικά για να εκπαιδεύσει ένα multilayer perceptron, ο αλγόριθμος των διανυσμάτων υποστήριξης είναι πιο γενικής φύσης και έχει ευρεία εφαρμοσιμότητα.

4 ΚΕΦΑΛΑΙΟ – ΔΙΑΓΡΑΜΜΑΤΑ ΕΛΕΓΧΟΥ

4.1 Στατιστικός έλεγχος ποιότητας

4.1.1 Εισαγωγή στην ποιότητα

Η ποιότητα αποτελεί έναν από τους βασικότερους παράγοντες για να επιλέξει κάποιος ένα προϊόν ή μια υπηρεσία. Η ποιότητα καθορίζει ακόμα και τις αποφάσεις εταιριών. Συνεπώς η ποιότητα οδηγεί σε εμπορική επιτυχία. Κάθε προϊόν έχει κάποια χαρακτηριστικά τα οποία καθορίζουν την ποιότητα. Αυτά τα κρίσιμα για την ποιότητα χαρακτηριστικά είναι πολλών ειδών όπως φυσικά, χημικά, ηλεκτρικά μεγέθη (μήκος, βάρος, τάση, ιξώδες), έκφραση αισθήσεων (χρώμα, εμφάνιση, γεύση), λειτουργικά στοιχεία (στιβαρότητα, αξιοπιστία, επισκευασιμότητα) και συμμόρφωση (κανονισμοί ασφάλειας τροφίμων, κανόνες καλής παραγωγής και πρότυπα).

Ποιότητα σημαίνει προσαρμογή στην χρήση. Υπάρχουν δύο γενικές απόψεις σχετικά με την ποιότητα: σχεδιασμός της ποιότητας και προσαρμογή της ποιότητας. Όλα τα προϊόντα και οι υπηρεσίες παράγονται σε διάφορα επίπεδα και σε διαφορετικό βαθμό ποιότητας. Η προσαρμογή της ποιότητας σχετίζεται με το πόσο καλά προσαρμόζεται το προϊόν στις προδιαγραφές με τις οποίες σχεδιάστηκε. Η ποιότητα είναι αντιστρόφως ανάλογη με την διασπορά. Αυτό σημαίνει ότι όταν η διασπορά ενός σημαντικού χαρακτηριστικού του προϊόντος μειώνεται τότε αυξάνεται η ποιότητα του προϊόντος.

4.1.2 Εισαγωγή στον στατιστικό έλεγχο διεργασιών (ΣΕΔ)

Ο στατιστικός έλεγχος διεργασιών είναι μια συλλογή από εργαλεία που χρησιμοποιούνται για τον εντοπισμό των αιτιών που ευθύνονται για την μεταβλητότητα μιας παραγωγικής διαδικασίας. Τα εργαλεία αυτά παρέχουν την δυνατότητα προσδιορισμού της «ικανότητας» μιας διαδικασίας να λειτουργεί και να παράγει σύμφωνα με τις απαιτούμενες προδιαγραφές. Η μεθοδολογία του ΣΕΔ παρέχει την απαραίτητη υποδομή για δειγματοληψία, ποιοτικό έλεγχο και ανάλυση και παράλληλα οι πληροφορίες που συλλέγονται χρησιμοποιούνται για την βελτίωση της παραγωγικής διαδικασίας. Με την μεθοδολογία του ΣΕΔ μπορεί εύκολα να εντοπιστεί η ύπαρξη ειδικών αιτιών που προκαλούν μεταβλητότητα σε μια διαδικασία.

- Τα επτά εργαλεία του ΣΕΔ

i. Ιστόγραμμα (histogram)/ φυλλόγραμμα (stem-and-leaf plot)
ii. Φύλλο ελέγχου (check sheet)
iii. Διάγραμμα Pareto (Pareto chart)
iv. Διάγραμμα αιτίας-αποτελέσματος (cause and effect diagram)
v. Διάγραμμα συγκέντρωσης ατελειών (defect concentration diagram)

vi. Διάγραμμα συσχετισμού (scatter diagram)
vii. Διάγραμμα ελέγχου (control chart)

Τα παραπάνω εργαλεία αποτελούν ένα σημαντικό τμήμα του ΣΕΔ με κυριότερο και πιο εύχρηστο τα διαγράμματα ελέγχου που αποτελούν ένα πανίσχυρο εργαλείο για παρακολούθηση σε πραγματικό χρόνο της ποιότητας παραγωγής και της πορείας των παραγωγικών διεργασιών. Με την συστηματική χρήση διαγραμμάτων ελέγχου επιτυγχάνεται μείωση της μεταβλητότητας και συνεπώς δημιουργία και διατήρηση σταθερών και ικανών παραγωγικών διαδικασιών.

- Το πρόβλημα του ΣΕΔ

Σε κάθε παραγωγική διεργασία, ανεξάρτητα από το πόσο καλά σχεδιασμένη είναι και το πόσο προσεκτικά επιβλέπεται και συντηρείται, θα υπάρχει μια μορφή φυσικής μεταβλητότητας. Αυτή η φυσική μεταβλητότητα είναι το αποτέλεσμα πολλών μικρών αιτιών οι οποίες αναφέρονται ως κοινές ή τυχαίες αιτίες μεταβλητότητας (common or chance causes of variation). Η φυσική μεταβλητότητα είναι συνήθως μικρή σε μέγεθος και δεν μπορεί να αποδοθεί σε ελέγξιμους παράγοντες. Μια διεργασία η οποία λειτουργεί μόνο με την παρουσία φυσικής μεταβλητότητας λέγεται εντός ελέγχου διεργασία (in control process) ή ότι λειτουργεί σε ευσταθή κατάσταση (stable state). Οι μορφές μεταβλητότητας που αφορούν την συστηματική αλλαγή στο επίπεδο κάποιου ή κάποιων παραγόντων που καθορίζουν την ποιότητα του προϊόντος οφείλονται συνήθως σε λανθασμένα ρυθμισμένες μηχανές, λάθη των χειριστών των μηχανημάτων, κακής ποιότητας ή ελαττωματική πρώτη ύλη. Η μεταβλητότητα που οφείλεται σε αυτούς τους λόγους είναι σε μέγεθος πολύ μεγαλύτερη της φυσικής και η παρουσία της οδηγεί συνήθως σε μη αποδεκτά επίπεδα λειτουργίας της παραγωγικής διαδικασίας. Αυτή η μεταβλητότητα καλείται ειδική μεταβλητότητα και οι αιτίες που οδηγούν σε αυτήν ονομάζονται ειδικές ή προσδιορισμένες αιτίες μεταβλητότητας (special or assignable causes of variation). Μια διεργασία η οποία λειτουργεί με την παρουσία ειδικής μεταβλητότητας καλείται εκτός ελέγχου διεργασία (out of control process) ή ότι λειτουργεί σε ασταθή κατάσταση (unstable state).

Το κύριο αντικείμενο του ΣΕΔ είναι η έγκαιρη ανίχνευση της εμφάνισης ειδικών αιτιών μεταβλητότητας σε μια διεργασία έτσι ώστε να προχωρήσουμε σε έρευνα και να προβούμε στις απαραίτητες διορθωτικές ενέργειες προτού κατασκευαστούν προϊόντα που δεν πληρούν τις προδιαγραφές. Τα διαγράμματα ελέγχου είναι μια τεχνική που χρησιμοποιείται για την ανίχνευση σε πραγματικό χρόνο της εμφάνισης ειδικών αιτιών μεταβλητότητας σε μια διεργασία (on-line process-monitoring).

4.1.3 Αποτελεσματικότητα ΣΕΔ

Για να είναι αποτελεσματικός ο Στατιστικός Έλεγχος Διεργασιών θα πρέπει να συνοδεύεται από ένα εκτός-ελέγχου πρόγραμμα δράσης (out-of-control action plan,

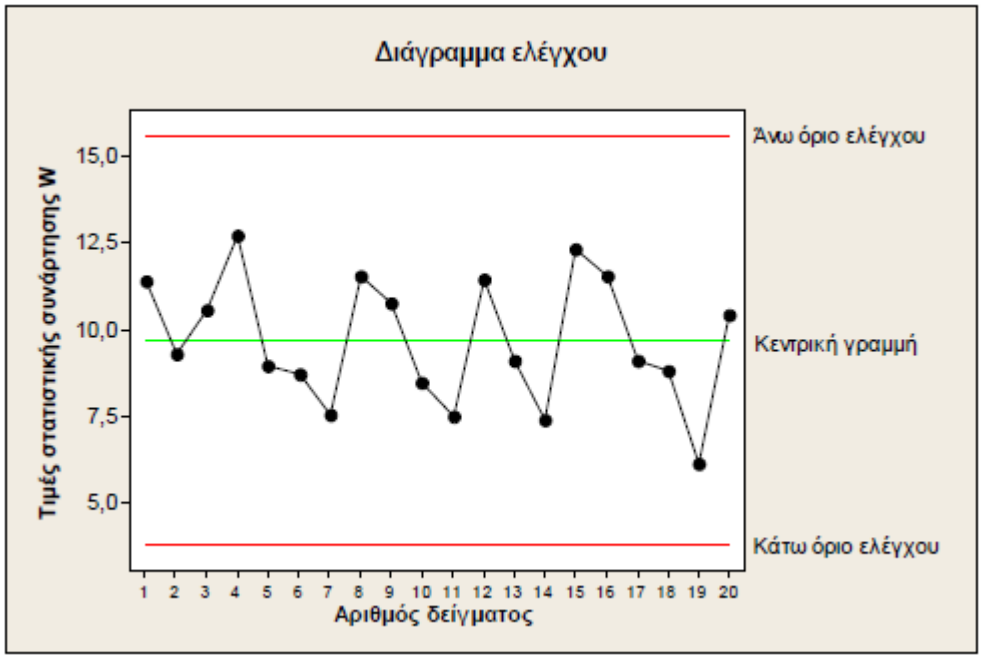
OCAP) το οποίο θα ενεργοποιείται κάθε φορά που το διάγραμμα ελέγχου παρέχει ενδείξεις εμφάνισης ειδικών αιτιών μεταβλητότητας στην διεργασία. Προκειμένου να αποφευχθεί η συσχέτιση των διαγραμμάτων ελέγχου μόνο με την περίπτωση που οι τιμές του ποιοτικού χαρακτηριστικού περιγράφονται από μια συνεχή τυχαία μεταβλητή (continuous variable), είναι απαραίτητο να αναφερθούν οι έννοιες μη συμμορφούμενο ή ελαττωματικό προϊόν (nonconforming or defective) και ο αριθμός ατελειών (defects or nonconformities). Ελαττωματικό προϊόν καλείται το προϊόν για το οποίο ένα ποιοτικό χαρακτηριστικό του έχει τιμή η οποία βρίσκεται εκτός των ορίων των προδιαγραφών. Στις περιπτώσεις τέτοιων προϊόντων κατασκευάζονται διαγράμματα ελέγχου για τον αριθμό των ελαττωμάτων, τα οποία ως ποιοτικά χαρακτηριστικά δεν μετρώνται σε μια συνεχή κλίμακα, αλλά παίρνουν αριθμήσιμο πλήθος τιμών και περιγράφονται με διακριτές τυχαίες μεταβλητές (discrete variables).

Σύμφωνα με τα παραπάνω, διακρίνονται δύο βασικές κατηγορίες διαγραμμάτων ελέγχου ανάλογα με το είδος της μεταβλητής που περιγράφει ένα ποιοτικό χαρακτηριστικό του προϊόντος:

i. Διαγράμματα ελέγχου για συνεχή χαρακτηριστικά-μεταβλητές (control charts for variables)
ii. Διαγράμματα ελέγχου για διακριτά χαρακτηριστικά-ιδιότητες (control charts for attributes)

4.1.4 Βασικές αρχές διαγραμμάτων ελέγχου

Η γενική θεωρία των διαγραμμάτων ελέγχου εισήχθηκε αρχικά από τον Dr. Walter S. Shewhart στα εργαστήρια της τηλεφωνικής εταιρίας Bell (1920) και για αυτόν τον λόγο ονομάζονται συχνά διαγράμματα Shewhart. Στο σχήμα που ακολουθεί παρουσιάζεται ένα τυπικό διάγραμμα ελέγχου. Πρόκειται για μια γραφική παράσταση ενός ποιοτικού χαρακτηριστικού σε συνάρτηση με τον χρόνο. Η κεντρική γραμμή παριστά την μέση τιμή του ποιοτικού χαρακτηριστικού που αντιστοιχεί στην υπό έλεγχο κατάσταση. Το άνω όριο ελέγχου και το κάτω όριο ελέγχου επιλέγονται έτσι ώστε αν η διαδικασία είναι υπό έλεγχο, τότε σχεδόν όλα τα σημεία του δείγματος βρίσκονται μεταξύ των δύο αυτών γραμμών. Καθώς τα σημεία κατανέμονται εντός των ορίων ελέγχου η διαδικασία είναι υπό έλεγχο και δεν απαιτείται καμία ενέργεια. Αν κάποιο σημείο βρεθεί εκτός των ορίων ελέγχου τότε η διαδικασία είναι εκτός ελέγχου και απαιτούνται ενέργειες για την εύρεση των αιτιών που προκαλούν το πρόβλημα. Τα σημεία του διαγράμματος συνδέονται με μια ευθεία γραμμή για να απεικονίζεται η εξέλιξη της διαδικασίας στον χρόνο. Ακόμα και αν όλα τα σημεία βρίσκονται εντός των ορίων ελέγχου αλλά τοποθετημένα με συστηματικό, μη τυχαίο τρόπο, τότε η διαδικασία είναι εκτός ελέγχου. Αν η διαδικασία είναι εντός ελέγχου τότε όλα τα σημεία του δείγματος πρέπει να κατανέμονται εντός των ορίων ελέγχου με τυχαίο τρόπο.



Εικόνα 1. - Ένα τυπικό διάγραμμα ελέγχου

Είναι αξιοσημείωτη η σχέση μεταξύ των διαγραμμάτων ελέγχου και του ελέγχου υποθέσεων. Αυτό γίνεται κατανοητό υποθέτοντας ότι η κεντρική γραμμή του σχήματος είναι ο δειγματικός μέσος $\bar{X}$. Αν αυτή η γραμμή είναι εντός των ορίων ελέγχου τότε η διαδικασία του μέσου είναι υπό έλεγχο και ισούται με μια τιμή μ_0. Αν ο $\bar{X}$ βρίσκεται εκτός των ορίων ελέγχου τότε η διαδικασία του μέσου είναι εκτός ελέγχου και ίση με μια τιμή $\mu_1 \neq \mu_0$. Δηλαδή το διάγραμμα ελέγχου αποτελεί έναν έλεγχο της υπόθεσης ότι η διαδικασία είναι σε κατάσταση ελέγχου. Υπάρχουν, όμως, και αρκετές διαφορές μεταξύ των διαγραμμάτων ελέγχου και των ελέγχων υποθέσεων.

Παρακάτω περιγράφεται ένα γενικό μοντέλο για το διάγραμμα ελέγχου:
Έστω W το στατιστικό ενός δείγματος που μετρά κάποιο ποιοτικό χαρακτηριστικό που μας ενδιαφέρει. Έστω ότι ο μέσος και η τυπική απόκλιση του W είναι μ_W και σ_W αντίστοιχα. Τότε, τα όρια ελέγχου και η κεντρική γραμμή θα είναι:

$$UCL = \mu_W + L\sigma_W$$
$$Κεντρική\ Γραμμή = \mu_W$$
$$LCL = \mu_W - L\sigma_W$$

όπου L είναι η απόσταση των ορίων ελέγχου από την κεντρική γραμμή.

- Επιλογή ορίων ελέγχου

Ο ορισμός των ορίων ελέγχου είναι από τις κρισιμότερες αποφάσεις κατά την κατασκευή ενός διαγράμματος ελέγχου. Απομακρύνοντας τα όρια ελέγχου από την

κεντρική γραμμή μειώνουμε το σφάλμα τύπου Ι, την ύπαρξη δηλαδή ενός σημείου εκτός ελέγχου που στην πραγματικότητα είναι εντός ελέγχου. Αυξάνεται, όμως, η πιθανότητα σφάλματος τύπου ΙΙ, δηλαδή τα σημεία φαίνεται να είναι εντός ελέγχου ενώ στην πραγματικότητα είναι εκτός. Μετακινώντας τα όρια ελέγχου κοντά στην κεντρική γραμμή αυξάνεται το σφάλμα τύπου Ι και μειώνεται το σφάλμα τύπου ΙΙ.

- Προειδοποιητικά όρια διαγράμματος ελέγχου

Μερικές φορές χρησιμοποιούνται δύο ζεύγη ορίων. Τα εξωτερικά είναι τα 3-σ όρια ελέγχου, κι όταν ένα σημείο βρεθεί εκτός αυτών, χρειάζεται διερεύνηση των αιτιών του προβλήματος και ενέργειες για την βελτίωση της παραγωγικής διαδικασίας. Τα εσωτερικά βρίσκονται σε απόσταση 2-σίγμα από την κεντρική γραμμή και ονομάζονται προειδοποιητικά όρια.

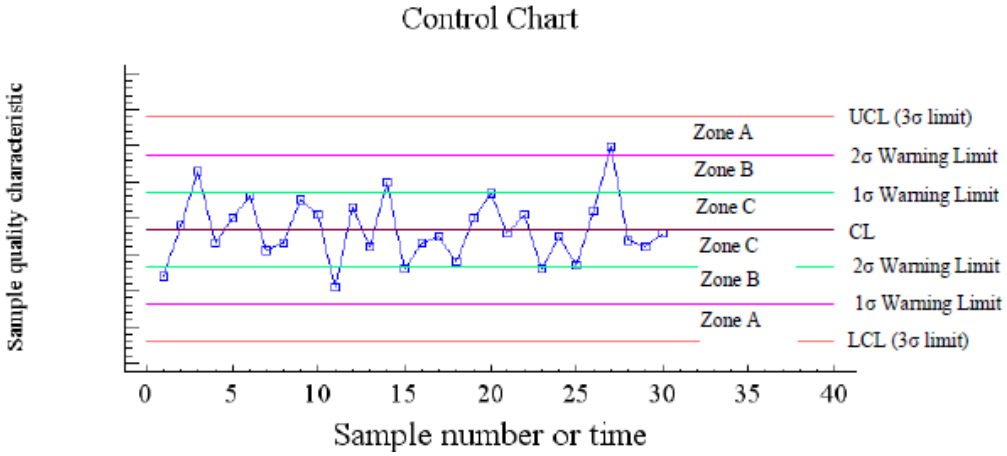
Αν ένα ή περισσότερα σημεία του δείγματος βρίσκονται στην περιοχή μεταξύ των προειδοποιητικών ορίων και των ορίων ελέγχου ή πολύ κοντά στα προειδοποιητικά όρια, υποπτευόμαστε ότι η διαδικασία δεν λειτουργεί σωστά. Η λύση είναι να αυξήσουμε τη συχνότητα ή και το μέγεθος των δειγμάτων ώστε να λαμβάνεται γρηγορότερα περισσότερη πληροφορία για την διαδικασία. Η χρήση προειδοποιητικών ορίων αυξάνει την «ευαισθησία» του διαγράμματος ελέγχου διότι εντοπίζονται τυχόν μετατοπίσεις του μέσου της διαδικασίας γρηγορότερα. Επομένως, αυξάνεται η πιθανότητα λανθασμένων σημάτων.

- Κανόνες ευαισθητοποίησης για τα διαγράμματα ελέγχου-Run Tests

Ένα διάγραμμα ελέγχου υποδεικνύει μια εκτός ελέγχου διαδικασία όταν ένα ή περισσότερα σημεία του δείγματος βρεθούν εκτός των ορίων ελέγχου ή όταν εμφανίζονται στο διάγραμμα ειδικές, μη τυχαίες ακολουθίες σημείων, τα patterns. Για τον εντοπισμό ασυνήθιστων, μη τυχαίων ακολουθιών σημείων σε ένα διάγραμμα χρησιμοποιούνται κάποιοι «κανόνες» (run tests) που περιγράφουν ενδεχόμενα που σχετίζονται με τηην εμφάνιση τέτοιων «ειδικών» ακολουθιών σημείων (patterns). Οι σημαντικότεροι κανόνες για την ευαισθητοποίηση ενός διαγράμματος ελέγχου είναι οι ακόλουθοι:

i. Ένα ή περισσότερα σημεία εκτός των ορίων ελέγχου.
ii. Δύο από τρία συνεχόμενα σημεία στην Ζώνη Α.
iii. Τέσσερα από τα πέντε συνεχόμενα σημεία πέραν της ζώνης C.
iv. Οκτώ συνεχόμενα σημεία πάνω ή κάτω από την κεντρική γραμμή.
v. Έξι συνεχόμενα σημεία σε αύξουσα ή φθίνουσα διάταξη.
vi. Δεκατέσσερα συνεχόμενα σημεία σε εναλλασσόμενη μορφή.
vii. Ένα ή περισσότερα σημεία κοντά σε οποιαδήποτε όρια.
viii. Δεκαπέντε συνεχόμενα σημεία στην ολική Ζώνη C.
ix. Οποιαδήποτε ασυνήθιστη ή μη τυχαία ακολουθία σημείων.

Οι τέσσερις πρώτοι κανόνες είναι γνωστοί ως Western Electric Rules.

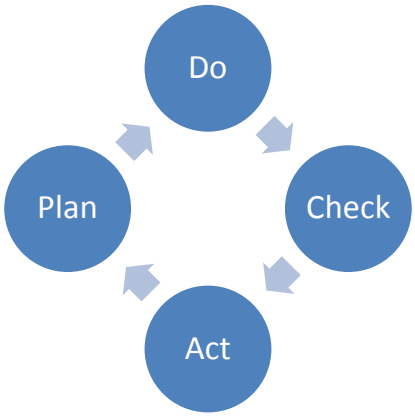


Εικόνα 2.- Ο διαχωρισμός των διαγραμμάτων ελέγχου σε ζώνες

4.2 Ιστορική Αναδρομή του Στατιστικού Ελέγχου Διεργασιών (ΣΕΔ)

4.2.1 Ο Deming και ο ΣΕΔ

Η φιλοσοφία του Deming υποστηρίζει την εφαρμογή ενός «Στατιστικο - Ποιοτικού» μοντέλου διοίκησης, δηλαδή μιας διοίκησης βασιζόμενης στην ευρεία χρήση στατιστικών μεθόδων για την λήψη αποφάσεων με σκοπό την συνεχή ποιοτική βελτίωση της παραγωγής και των υπηρεσιών. Αυτό απαιτεί την ελαχιστοποίηση της μεταβλητότητας (variability), το οποίο οδηγεί σε λιγότερες δαπάνες, λάθη, κόστος, άσκοπη εργασία καθώς επίσης και σε υψηλότερο βαθμό αξιοπιστίας και στην απόκτηση καλής φήμης και μεριδίου αγοράς. Ο Edward Deming, ο οποίος θεωρείται από πολλούς ως ο πατέρας της Διοίκησης Ολικής Ποιότητας, είναι υπεύθυνος για την ποιοτική επανάσταση που επιτεύχθηκε στην Ιαπωνία μετά το Β' Παγκόσμιο Πόλεμο και κατέληξε στην κυριαρχία της Ιαπωνικής βιομηχανίας. Στόχος της Διοίκησης Ολικής Ποιότητας είναι η συνεχής βελτίωση στην ποιότητα εκτέλεσης όλων των διεργασιών, προϊόντων και υπηρεσιών σε μια επιχείρηση. Αυτό επιτυγχάνεται με τέσσερα βήματα που επαναλαμβάνονται το ένα μετά το άλλο, και μετά το τελευταίο βήμα η διαδικασία ξεκινά από την αρχή (Koukouvinos Ch, 2008).Τα βήματα αυτά αποτελούν τον τροχό του Deming ο οποίος απεικονίζεται παρακάτω:



Εικόνα 1 - Ο τροχός του Deming

Οι κανόνες του Deming για την διεύθυνση μιας εταιρίας περιγράφονται ως εξής:

- Εφαρμογή ενός προγράμματος που αποβλέπει στην συνεχή βελτίωση των προϊόντων και της προσφερόμενης υπηρεσίας προς τον καταναλωτή.
- Προσπάθεια για μακράς διάρκειας οφέλη και όχι για μικρής διάρκειας κέρδη.
- Ανεξαρτητοποίηση της παραγωγής από τον τελικό συνολικό ποιοτικό έλεγχο.
- Εξάλειψη των φραγμών, του μυστικισμού και του συναγωνισμού μεταξύ των τμημάτων της ίδιας εταιρείας.
- Εκπαίδευση στις στατιστικές μεθόδους για όλο το προσωπικό.

Τα 14 βασικά σημεία της φιλοσοφίας του Deming παρουσιάζονται στον ακόλουθο πίνακα.

Κανόνας 1	Δημιούργησε το κλίμα για ένα σταθερό πρόγραμμα που να αποβλέπει στη συνεχή βελτίωση των προϊόντων και των υπηρεσιών.
Κανόνας 2	Υιοθέτησε τη νέα φιλοσοφία της απόρριψης της κακής δουλειάς, των ελαττωματικών και των μη ικανοποιητικών υπηρεσιών.
Κανόνας 3	Μην βασίζεστε στον τελικό έλεγχο όλης της παραγωγής για την επίτευξη υψηλής ποιότητας.
Κανόνας 4	Σταματήστε τη συνήθεια αγοράς προμηθειών με τη χαμηλότερη τιμή.
Κανόνας 5	Προσπαθήστε να βελτιώνεται χωρίς διακοπή όλα τα συστήματα παραγωγής και υπηρεσιών.
Κανόνας 6	Θεσπίστε ένα πρόγραμμα σύγχρονης εκπαίδευσης και εφαρμόστε το σε όλους τους εργαζόμενους.
Κανόνας 7	Υιοθετήστε σύγχρονους τρόπους επίβλεψης.
Κανόνας 8	Εξαλείψτε το φόβο. Μη διστάζετε να ρωτάτε και να αναφέρετε προβλήματα.
Κανόνας 9	Εξαφανίστε τα εμπόδια μεταξύ των τμημάτων της επιχείρησης. Η από κοινού δουλειά μεταξύ των διαφόρων τμημάτων της επιχείρησης είναι απαραίτητο στοιχείο για την ανάπτυξη της ποιότητας.
Κανόνας 10	Εξαλείψτε τους στόχους και τα συνθήματα για μηδέν ελαττωματικά προϊόντα.
Κανόνας 11	Εξαφανίστε τους αριθμητικούς στόχους.
Κανόνας 12	Εξαλείψτε τα εμπόδια που αφαιρούν από τον εργαζόμενο το δικαίωμα

	να νιώθει υπερήφανος για την δουλειά του.
Κανόνας 13	Εφαρμόστε ένα πρόγραμμα συνεχούς ενημέρωσης και εκπαίδευσης για όλους τους εργαζομένους.
Κανόνας 14	Δημιουργήστε ένα ανώτατο κλιμάκιο διοίκησης το οποίο θα εργαστεί σθεναρά για την επίτευξη των πρώτων 13 κανόνων.

Πίνακας 2 -. Οι 14 κανόνες του Deming

Οι κανόνες του Deming συμπεριλαμβάνονται σε τρεις βασικές αρχές, τις οποίες ο B.L. Joiner θέτει στις κορυφές ενός ισόπλευρου τριγώνου:

- Επιμονή για ποιότητα: Ο προφανέστερος τρόπος για την βελτίωση της ποιότητας των προϊόντων είναι η συνεχής προσπάθεια βελτίωσης των μέσων παραγωγής και προγραμμάτων.
- Όλοι σε μια ομάδα: Πρέπει να δημιουργηθεί ένα αίσθημα σε όλο το προσωπικό μιας εταιρίας ότι ανήκουν στην ίδια ομάδα και έχουν κοινούς στόχους και προσδοκίες.
- Χρήση της επιστημονικής μεθόδου: Η χρήση της επιστημονικής μεθόδου είναι ο καλύτερος και πιο σίγουρος τρόπος επίτευξης της ποιοτικής βελτίωσης. Στατιστικές μέθοδοι ανάλυσης, χρησιμοποιώντας πραγματικά δεδομένα, μπορούν να βοηθήσουν στην λήψη αποφάσεων για την βελτίωση της ποιότητας.

Σύμφωνα με τον Deming, η γνώση και η εφαρμογή στατιστικών θεωριών και μεθόδων είναι σημαντική και αναγκαία, επειδή η στατιστική ανάλυση των δεδομένων χρησιμοποιείται ως μηχανισμός αξιοποίησης και εκμετάλλευσης των πληροφοριών που προκύπτουν πριν και κατά την διάρκεια της παραγωγής, ώστε να επιτευχθεί η πρόβλεψη, αναγνώριση και διόρθωση λαθών με σκοπό την συνεχή ποιοτική βελτίωση. Η κυριότερη συνεισφορά του είναι ότι τόνισε την ανάγκη να μετακινηθεί προς τα πίσω ο ποιοτικός έλεγχος, δηλαδή από την τελική επιθεώρηση στον κατάλληλο έλεγχο της διεργασίας. Σε τρεις λέξεις η φιλοσοφία του Deming συνοψίζεται: «Ελαττώστε τη μεταβλητότητα». Για τον έλεγχο και την μείωση της μεταβλητότητας κατά την διαδικασία της παραγωγής (on-line quality control) ο Deming πρότεινε το Στατιστικό Έλεγχο Διεργασίας ή ΣΕΔ (Statistical Process Control).

Ο ΣΕΔ είναι μια τεχνική που κατατάσσεται σε υψηλά επίπεδα μεταξύ των προγραμμάτων που σχετίζονται με την βελτίωση της ποιότητας και ως κύριο εργαλείο του έχει τα διαγράμματα που δημιούργησε ο Shewhart. Η ελάττωση της μεταβλητότητας υπήρξε ουσιαστική επιδίωξη του Walter Shewhart, ο οποίος δημιούργησε το διάγραμμα ελέγχου (control chart). Το έργο του στην δεκαετία του 1920 εστιάστηκε στην μείωση της μεταβλητότητας που αφορούσε στην λειτουργία των τηλεφώνων στα εργαστήρια του Bell. Ο ΣΕΔ περιλαμβάνει ένα επιστημονικό πεδίο που είναι η εκτέλεση απαραίτητων υπολογισμών ώστε να μπορούν να εισαχθούν στο διάγραμμα γραμμές σχετιζόμενες με την απόδοση και τα όρια ελέγχου.

Τα όρια ελέγχου δεν είναι όρια προδιαγραφών, δεν εξαρτώνται από τυχόν απαιτήσεις αλλά εξαρτώνται μόνο από το πώς λειτουργεί η διεργασία την συγκεκριμένη χρονική περίοδο.

Εφόσον οι μετρήσεις των δειγμάτων, όσον αφορά την απόδοση, παραμένουν ανάμεσα στο άνω όριο ελέγχου (upper control limit) και το κάτω όριο ελέγχου (lower control limit), η διεργασία είναι υπό έλεγχο. Αντίθετα, η μη τυχαία συμπεριφορά ή οι αποκλίσεις εκτός των ορίων απαιτούν άμεσες διορθωτικές αλλαγές στην διεργασία, ώστε να βρεθεί υπό έλεγχο, δηλαδή σε μια σταθερή κατάσταση. Αυτή η κατάσταση ονομάζεται κατάσταση στατιστικού ελέγχου και η μεταβλητότητα είναι ελέγξιμη και προβλέψιμη. Ο στατιστικός έλεγχος διεργασίας, εκτός από το να υπολογίζει την απόδοση και να προσδιορίζει το αν συμμορφώνεται ή όχι με τις απαιτήσεις του στατιστικού ελέγχου, επιδιώκει να καθοδηγήσει ενέργειες επί της διεργασίας (on-line), στον κατάλληλο χρόνο ώστε η διασπορά της διεργασίας να ελαχιστοποιηθεί. Ο χρόνος δράσης, το είδος των ενεργειών και η ευθύνη για αυτές, εξαρτώνται από το αν τα αίτια της μεταβλητότητας είναι ελεγχόμενα (κοινά) ή μη ελεγχόμενα (ειδικά).

Σύμφωνα με τον Deming, τα κοινά αίτια αναφέρονται στις διαφορετικές πηγές διασποράς μιας διεργασίας που βρίσκεται υπό στατιστικό έλεγχο. Αυτά μπορεί να είναι οι μη ελεγχόμενες περιβαλλοντικές συνθήκες, η μεταβλητότητα των αγορασμένων υλικών και άλλα αίτια. Η ανάλυση των κοινών αιτιών της διασποράς απαιτεί δράση επί του συστήματος. Η παραβίαση των ορίων ενός διαγράμματος ή η παρουσία ενός συγκεκριμένου διαγράμματος εντός ορίων, είναι ένδειξη ύπαρξης ειδικών αιτιών παρέκκλισης, όπως αλλαγές χειριστού βάρδιας, απώλειες λόγω καταστροφής των μηχανημάτων κλπ. Η ανακάλυψη και η απομάκρυνσή τους απαιτούν επί τόπου ενέργειες από κάποιον που συνδέεται άμεσα με την παραγωγική διαδικασία.

4.2.2 Από τον Deming στον Taguchi

Μία από τις βασικές συνεισφορές του Deming ήταν το ότι έπεισε τον κόσμο να στρέψει τις προσπάθειές του για βελτίωση της ποιότητας από την μαζική εποπτεία στον έλεγχο διεργασίας, μέσω του Στατιστικού Ελέγχου Διεργασίας. Η συνεισφορά του Taguchi ήταν το ότι έκανε ένα βήμα πίσω, δηλαδή από την παραγωγή στο σχεδιασμό. Το στάδιο του σχεδιασμού είναι το εκτός-διεργασίας στάδιο ελέγχου ποιότητας (off-line quality control). Την δεκαετία του 1980, ο Genichi Taguchi εισήγαγε νέες ιδέες σχετικά με την βελτίωση της ποιότητας στις Ηνωμένες Πολιτείες. Παρουσίασε μία καινοτομική προσέγγιση παραμετρικών σχεδιασμών για την μείωση της διακύμανσης στα προϊόντα και τις διαδικασίες. Οι μέθοδοί του και η φιλοσοφία του δημιούργησαν σημαντικό ενδιαφέρον μεταξύ των μηχανικών ποιότητας και των στατιστικολόγων. Η μεθοδολογία του χρησιμοποιήθηκε από τα AT&T Bell εργαστήρια, την εταιρία Ford motor company και την Xerox καθώς και άλλους παράγοντες της βιομηχανίας στην Αμερική.

Ο Taguchi έχοντας σαν αφετηρία τον τρίτο κανόνα του Deming για την διοίκηση, επινόησε μια τεχνική βελτίωσης που χρησιμοποιεί τις μεθόδους του Στατιστικού Σχεδιασμού Πειραμάτων (Statistical Design of Experiments-SDE), για τον αποτελεσματικό χαρακτηρισμό ενός προϊόντος ή των μέσων παραγωγής σε συνδυασμό με την στατιστική ανάλυση της διασποράς τους, επιδιώκοντας την ελαχιστοποίηση της μεταβλητότητας με το χαμηλότερο κόστος. Κατά την εφαρμογή του ΣΕΔ κατά την διάρκεια της παραγωγής (on-line quality control) αλλάζουμε κάθε φορά έναν παράγοντα, ο οποίος προκαλεί τη ζημιά, ενώ η μέθοδος του Taguchi με την βοήθεια των στατιστικών σχεδιασμών εφαρμόζει την αρχή της αλλαγής πολλών παραγόντων κάθε φορά, το οποίο κοστίζει λιγότερο και δίνει πιο αξιόπιστα αποτελέσματα. Για τον έλεγχο και την μείωση της μεταβλητότητας καθώς και την οικοδόμηση της ποιότητας σε μια πρώιμη φάση του προϊόντος ή της ανάπτυξης των μέσων παραγωγής, οι τεχνικές του Taguchi για τον εκτός-διεργασίας έλεγχο ποιότητας προσελκύουν ιδιαίτερο ενδιαφέρον. Για να διασφαλιστεί η ποιότητα ο ΣΕΔ μπορεί να εφαρμοστεί σε μεταγενέστερη φάση ώστε να ολοκληρωθεί ο κύκλος του ελέγχου ποιότητας, ο οποίος ονομάζεται ολικός έλεγχος ποιότητας. Για την αποτελεσματική χρήση του Στατιστικού Ελέγχου Ποιότητας είναι απαραίτητη η ενσωμάτωσή του στην διοίκηση της επιχείρησης που έχει ως στόχο τη συνεχή βελτίωση της ποιότητας, μέσα από την Διοίκηση Ολικής Ποιότητας (Total Quality Management/ Total Quality Assurance).

Ο Στατιστικός Έλεγχος Ποιότητας αποτέλεσε το πρώτο βήμα για την Διοίκηση Ολικής Ποιότητας. Στον ακόλουθο πίνακα παρουσιάζονται τα κυριότερα ιστορικά σημεία στην ανάπτυξη της ποιότητας από το 1900 μέχρι και σήμερα.

1901	Τα πρώτα εργαστήρια προτύπων (standards) ιδρύονται στην Μ. Βρετανία.
1907	Η AT & Bell Laboratories αρχίζει τη συστηματική επιθεώρηση και έλεγχο προϊόντων και υλικών.
1919	Η Ένωση Τεχνικών Επιθεωρητών ιδρύεται στην Αγγλία, η οποία αργότερα μετονομάζεται σε Ινστιτούτο Διασφάλισης της Ποιότητας
1920	Στα εργαστήρια της AT & Bell Laboratories ιδρύεται τμήμα ποιότητας.
1924	Ο W. A. Shewhart εισάγει τα διαγράμματα ελέγχου σε ένα Technical Report στην AT & Bell.
1928	Το δειγματοληπτικό u963 σχέδιο αποδοχής σωρού αναπτύσσεται από τους Dodge & Romig
1931	Ο W.A. Shewhart εκδίδει το περιοδικό Economic Control of Quality of Manufactured Product.
1932	Ο W.A. Shewhart δίνει διαλέξεις σε στατιστικές μεθόδους στην παραγωγή και στα διαγράμματα ελέγχου στο Πανεπιστήμιο του Λονδίνου.
1938	Ο W.E. Deming προσκαλεί τον Shewhart για σεμινάρια στα διαγράμματα ελέγχου στο U.S. Departement of Agriculture Graduate School.
1940	Το Υπουργείο Πολέμου των Η.Π.Α. εκδίδει έναν οδηγό για την ανάλυση δεδομένων με την χρήση των διαγραμμάτων ελέγχου.
1946	Ιδρύεται η American Society for Quality Control (ASQC).
1946	Ο W.E. Deming προσκαλείται στην Ιαπωνία για να δώσει σεμινάρια Στατιστικού Ποιοτικού Ελέγχου.
1948	Ο καθηγητής G. Taguchi αρχίζει την μελέτη των πειραματικών σχεδιασμών.

1950	Ο W.E. Deming αρχίζει την εκπαίδευση ανωτάτων στελεχών βιομηχανιών της Ιαπωνίας.
1950	Ο K. Ishikawa εισάγει το διάγραμμα αιτίου-αποτελέσματος (cause & effect diagram)
1954	Ο J.M. Juran προσκαλείται από την ιαπωνία να δώσει διαλέξεις σε θέματα διοίκησης και βελτίωσης της ποιότητας. Ο E.S. Page εισάγει το διάγραμμα ελέγχου CUSUM.
1959	Ο S. Roberts εισάγει το διάγραμμα ελέγχου EWMA. Αρχίζει να εκδίδεται το επιστημονικό περιοδικό Technometrics.
1960	Εισάγεται από τον K. Ishikawa η έννοια των κύκλων ποιότητας.
1969	Αρχίζει η έκδοση των περιοδικών Quality Progress και Journal of Quality Technology.
1975	Εκδίδονται τα πρώτα βιβλία σε σχεδιασμό πειραμάτων.
1989	Ξεκινά η έκδοση του περιοδικού Quality Engeneering. Η Motorola εισάγει την έννοια six-sigma.
1990	Σταδιακή αύξηση της ζήτησης στην βιομηχανία για πιστοποιητικά κατά ISO 9000.
1997	Η προσέγγιση six-sigma της Motorola υιοθετείται και από άλλες βιομηχανίες.
2000	Δεύτερη αναθεώρηση της σειράς προτύπων ISO 9000.
2003	Ενίσχυση της δέσμευσης, της υπευθυνότητας και της επίγνωσης των προμηθευτών με τις επιχειρήσεις σχετικά με την ποιότητα.
2005	Θεσπίζεται από τους Filho και Cezar η εφαρμογή διαδικασιών βελτίωσης του Συστήματος Διοίκησης Ποιότητας.
2006	Θεσπίζεται το μοντέλο αξιολόγησης προγράμματος στην τριτοβάθμια εκπαίδευση.

Πίνακας 3 - Ιστορική αναδρομή της ποιότητας

5 ΚΕΦΑΛΑΙΟ – ΒΑΣΙΚΑ ΔΙΑΓΡΑΜΜΑΤΑ ΕΛΕΓΧΟΥ

5.1 Διαγράμματα ελέγχου μεταβλητών

5.1.1 Εισαγωγή

Τα διαγράμματα ελέγχου μεταβλητών αναπτύχθηκαν το 1930 και εφαρμόζονται στην βιομηχανία και σε πολλές επιστημονικές περιοχές. Όταν το πρόβλημα αφορά μεταβλητή, πρέπει να ελέγξουμε τη μέση τιμή και την διασπορά του ποιοτικού χαρακτηριστικού. Ο έλεγχος για την μέση τιμή κατά την διάρκεια μια διαδικασίας γίνεται συνήθως με τα διαγράμματα ελέγχου για τον μέσο -$\bar{X}$ διαγράμματα. Ο έλεγχος για την διασπορά γίνεται είτε με τα διαγράμματα ελέγχου για την τυπική απόκλιση (S- διαγράμματα), είτε με τα διαγράμματα ελέγχου για το εύρος (R- διαγράμματα). Τα $\bar{X}$ και R διαγράμματα αποτελούν τα πιο χρήσιμα εργαλεία του Στατιστικού ελέγχου διεργασίας. (Koukouvinos Ch., 2008)

5.1.2 Διαγράμματα ελέγχου $\bar{X} - R$

Έστω ότι το χαρακτηριστικό X των προϊόντων που παράγονται ακολουθεί την κανονική κατανομή $N(\mu, \sigma^2)$ με μ, σ γνωστά. Αν $X_i = (X_{i1}, X_{i2}, \dots, X_{in}), i \geq 1$ είναι ένα τυχαίο δείγμα μεγέθους n από την X τότε ο δειγματικός μέσος:

$$\bar{X}_i = \frac{X_{i1} + X_{i2} + \cdots + X_{in}}{n}$$

ακολουθεί την κανονική κατανομή $N(\mu, \sigma^2/n)$ και είναι αμερόληπτη εκτιμήτρια της μέσης τιμής μ του χαρακτηριστικού X. Ο δειγματικός μέσος $\bar{X}_i$ παίρνει τιμές στο διάστημα

$$\left[\mu_{\bar{X}_i} - z_{\frac{a}{2}}\sigma_{\bar{X}_i}, \qquad \mu_{\bar{X}_i} + z_{\frac{a}{2}}\sigma_{\bar{X}_i}\right], \sigma_{\bar{X}_i} = \frac{\sigma}{\sqrt{n}}$$

με πιθανότητα 1-α.

Επομένως, χρησιμοποιώντας ένα διάγραμμα ελέγχου στο οποίο απεικονίζεται η τιμή του δειγματικού μέσου $\bar{X}_i$, τότε στα δείγματα που επιλέγουμε από την παραγωγή με

$$LCL_{\bar{X}} = \mu_{\bar{X}_i} - 3\sigma_{\bar{X}_i} = \mu - 3\frac{\sigma}{\sqrt{n}}$$
$$CL_{\bar{X}} = \mu_{\bar{X}_i} = \mu$$
$$UCL_{\bar{X}} = \mu_{\bar{X}_i} + 3\sigma_{\bar{X}_i} = \mu + 3\frac{\sigma}{\sqrt{n}}$$

και υποθέτοντας ότι η διακύμανση (ή γενικότερα η διασπορά) του χαρακτηριστικού X σε όλη την διαδικασία παραμένει σταθερή, τότε τα συμπεράσματα είναι τα παρακάτω:

i. Όταν τα σημεία του διαγράμματος βρίσκονται εντός των ορίων ελέγχου θεωρούμε ότι η διαδικασία είναι εντός ελέγχου και η μέση τιμή μ του χαρακτηριστικού X δεν έχει μετατοπιστεί και επομένως το 99.73% των σημείων του διαγράμματος βρίσκονται εντός των ορίων ελέγχου.
ii. Αν ένα σημείο βρίσκεται εκτός των ορίων ελέγχου, επειδή η πιθανότητα αυτού του ενδεχομένου είναι πολύ μικρή (0.0027), τότε θα υπάρχει ένδειξη ότι η διαδικασία είναι εκτός ελέγχου λόγω μετατόπισης της μέσης τιμής του χαρακτηριστικού X.

Οι τιμές των μ, σ είναι άγνωστες και πρέπει να εκτιμηθούν. Για να επιτευχθεί αυτό, επιλέγονται m=20 έως 25 ανεξάρτητα τυχαία δείγματα μεγέθους n=4 έως 6 το καθένα, υποθέτοντας ότι η επιλογή των δειγμάτων έγινε όταν η διεργασία ήταν εντός ελέγχου.

- Εκτίμηση του μ

Έστω $\bar{X}_i, \dots, \bar{X}_m$ οι δειγματικοί μέσοι των m δειγμάτων.
Θέτουμε

$$\bar{\bar{X}} = \frac{1}{m}\sum_{i=1}^{m} \bar{X}_i = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}$$

Τότε, ανεξάρτητα από την κατανομή του πληθυσμού, από το Κεντρικό Οριακό Θεώρημα ισχύει

$$E\left(\bar{\bar{X}}\right) = \mu$$
$$Var\left(\bar{\bar{X}}\right) = \frac{\sigma^2}{nm}$$

Η ποσότητα $\bar{\bar{X}}$ ακολουθεί την κανονική κατανομή $N(\mu, \sigma^2/nm)$ και χρησιμοποιείται ως εκτίμηση της ποσότητας μ. Δηλαδή $\hat{\mu} = \bar{\bar{X}}$.

- Εκτίμηση του σ

  - Μέθοδος R

Έστω $R_1, R_2, \dots, R_m$ τα εύρη m δειγμάτων.

$$R_i = \max\{X_{i1}, \dots, X_{in}\} - min\{X_{i1}, \dots, X_{in}\}, i = 1, \dots, m$$

$$\mu_{R_i} = E(R_i) = \sigma d_2$$
$$\sigma_{R_i} = \sqrt{Var(R_i)} = \sigma d_3$$

Θέτοντας

$$\bar{R} = \frac{R_1 + R_2 + \cdots + R_m}{m}$$

προκύπτει

$$E(\bar{R}) = \sigma d_2$$

Δηλαδή η ποσότητα $\bar{R}/_{d_2}$ είναι αμερόληπτη εκτιμήτρια της ποσότητας σ.

Δηλαδή $\hat{\sigma} = \bar{R}/_{d_2}$.

- Μέθοδος S

Έστω $S_i = \sqrt{{S_i}^2} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2}, 1 \le i \le m$

$$\mu_{S_i} = E(S_i) = \sigma c_4$$
$$\sigma_{S_i} = \sqrt{Var(S_i)} = \sigma\sqrt{1 - c_4^2}$$

όπου c_4 σταθερά που εξαρτάται από το μέγεθος n των δειγμάτων.
Θέτουμε

$$\bar{S} = \frac{S_1 + S_2 + \cdots + S_m}{m}$$

οπότε

$$E(\bar{S}) = \sigma c_4$$

Δηλαδή η ποσότητα $\bar{S}/_{c_4}$ είναι αμερόληπτη εκτιμήτρια της ποσότητας σ.

Δηλαδή $\hat{\sigma} = \bar{S}/_{c_4}$.

Ανάλογα με την εκτίμηση που χρησιμοποιούμε για την τυπική απόκλιση σ , προκύπτουν διαφορετικά διαγράμματα ελέγχου για την μέση τιμή και τη διασπορά. Τα διαγράμματα $\bar{X} - R$ και $\bar{X} - S$.
Τα όρια ελέγχου για το $\bar{X}$- διάγραμμα προκύπτουν από τις παρακάτω σχέσεις:

$$UCL = \bar{\bar{X}} + A_2\bar{R}$$
$$Center\ Line = \bar{\bar{X}}$$
$$UCL = \bar{\bar{X}} - A_2\bar{R}$$

Τα όρια ελέγχου για το R-διάγραμμα προκύπτουν από τις παρακάτω σχέσεις:

$$UCL = D_4\bar{R}$$
$$Center\ Line = \bar{R}$$
$$UCL = D_3\bar{R}$$

όπου

$$\bar{\bar{X}} = \frac{1}{m}\sum_{i=1}^{m}\bar{X}_i$$

$$\bar{X}_i = \frac{X_{i1} + X_{i2} + \cdots + X_{in}}{n} \; i = 1, \dots, m$$

$$\bar{R} = \frac{1}{m}\sum_{i=1}^{m} R_i$$

$$R_i = \max\{X_{i1}, \dots, X_{in}\} - min\{X_{i1}, \dots, X_{in}\}, i = 1, \dots, m$$

$$A_2 = \frac{3}{d_2\sqrt{n}}$$

$$D_3 = 1 - 3\frac{d_3}{d_2}$$

$$D_4 = 1 + 3\frac{d_3}{d_2}$$

Οι συντελεστές A_2, D_3, D_4 εξαρτώνται από το μέγεθος n του δείγματος.

- Κατασκευή $\bar{X} - R$ διαγραμμάτων ελέγχου

Βήματα:

i. Κατά την διάρκεια της κανονικής παραγωγής καταγράφουμε τις μετρήσεις m=20 δειγμάτων μεγέθους n=5. Αυτοί οι αριθμοί είναι σύμφωνοι με τον κανόνα που λέει ότι για μια αρχική μελέτη διεργασίας με μεταβλητά δεδομένα απαιτούνται τουλάχιστον 20 δείγματα μεγέθους 5, για να απεικονίζουμε επαρκώς την φυσιολογική μεταβλητότητα που υπάρχει στην διεργασία. Η συχνότητα του δείγματος εξαρτάται από τον όγκο της παραγωγής.

ii. Για καθένα από τα m δείγματα καταγράφουμε τον μέσο (αριθμητικός μέσος) $\bar{X}_i$ και το εύρος R_i , $1 \le i \le m$.

iii. Υπολογίζουμε το συνολικό μέσο όρο και το μέσο όρο του εύρους:

$$\bar{\bar{X}} = \frac{1}{m}\sum_{i=1}^{m} \bar{X}_i$$

$$\bar{R} = \frac{1}{m}\sum_{i=1}^{m} R_i$$

$$1 \le i \le m$$

iv. Υπολογίζουμε τα άνω και κάτω όρια ελέγχου για τα δύο διαγράμματα. Το εύρος ζώνης των τιμών που καλύπτονται από τα όρια ελέγχου στο $\bar{X}$-διάγραμμα είναι η προσέγγιση ενός εύρους ζώνης 6 τυπικών αποκλίσεων στον κανονικό πληθυσμό των μέσων όρων του δείγματος. Η κατανομή των ευρών του δείγματος δεν είναι κανονική αλλά ανισοβαρής, ιδιαίτερα στα μικρά δείγματα.

v. Σχεδιάζουμε τα διαγράμματα $\bar{X}$ και R. Αν όλες οι τιμές $\bar{X}_i$ και R_i (για κάθε δείγμα) βρίσκονται με τυχαίο τρόπο εντός των ορίων ελέγχου, τότε η διεργασία θεωρείται ότι βρίσκεται υπό στατιστικό έλεγχο. Διαφορετικά υπάρχουν ειδικά αίτια μεταβλητότητας τα οποία πρέπει να διερευνηθούν και να εξαλειφθούν.

5.1.3 Διαγράμματα ελέγχου $\bar{X} - S$

Τα $\bar{X} - S$ διαγράμματα ελέγχου χρησιμοποιούνται κυρίως όταν το μέγεθος n του δείγματος είναι σχετικά μεγάλο, n>10 ή 12. Η διαδικασία κατασκευής των διαγραμμάτων $\bar{X} - S$ είναι ίδια με αυτή για τα $\bar{X} - R$. Η διαφορά είναι ότι για κάθε δείγμα πρέπει να υπολογίζεται ο μέσος $\bar{X}$ και η τυπική απόκλιση S. Για την διασπορά του ποιοτικού χαρακτηριστικού X χρησιμοποιείται η στατιστική συνάρτηση

$$S_i = \sqrt{{S_i}^2} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2}\,, 1 \le i \le m,$$

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 c_4 σταθερά που εξαρτάται από το μέγεθος n των δειγμάτων.

Ένα διάγραμμα ελέγχου για την διασπορά του χαρακτηριστικού X μπορεί να βασιστεί σε ένα διάγραμμα των δειγματικών τυπικών αποκλίσεων S_i. Το μοντέλο με όρια 3-σ θα έχει την μορφή:

$$UCL_S = \mu_{S_i} + 3\sigma_{S_i} = \left(c_4 + 3\sqrt{1 - c_4^2}\right)\sigma$$

$$CL_S = \mu_{S_i} = c_4\sigma$$

$$LCL_S = \mu_{S_i} - 3\sigma_{S_i} = \left(c_4 - 3\sqrt{1 - c_4^2}\right)\sigma$$

Θέτοντας $B_5 = c_4 - 3\sqrt{1 - c_4^2}$ και $B_6 = c_4 + 3\sqrt{1 - c_4^2}$, τότε οι παράμετροι τροποποιούνται ως εξής:

$$UCL_S = B_6\sigma$$

$$CL_S = c_4\sigma$$

$$LCL_S = B_5\sigma$$

Όμως, η ποσότητα σ είναι άγνωστη, αλλά η εκτίμησή της είναι: $\hat{\sigma} = \bar{S}/c_4$. Δηλαδή τα όρια ελέγχου θα είναι:

$$UCL_S = \mu_{S_i} + 3\sigma_{S_i} = \bar{S} + 3\sigma\sqrt{1 - c_4^2} = \bar{S} + 3\frac{\bar{S}}{c_4}\sqrt{1 - c_4^2} = \left(1 + 3\frac{1}{c_4}\sqrt{1 - c_4^2}\right)\bar{S}$$
$$= B_4\bar{S}$$

$$CL_S = \mu_{S_i} = \bar{S}$$

$$LCL_S = \mu_{S_i} - 3\sigma_{S_i} = \bar{S} - 3\sigma\sqrt{1 - c_4^2} = \bar{S} - 3\frac{\bar{S}}{c_4}\sqrt{1 - c_4^2} = \left(1 - 3\frac{1}{c_4}\sqrt{1 - c_4^2}\right)\bar{S}$$
$$= B_3\bar{S}$$

Τα αντίστοιχα όρια ελέγχου για το $\bar{X}$-διάγραμμα χρησιμοποιώντας την εκτίμηση $\hat{\sigma} = \bar{S}/c_4$ μετασχηματίζονται ως εξής:

$$UCL_{\bar{X}} = \mu_{\overline{X_i}} + 3\sigma_{\overline{X_i}} = \bar{\bar{X}} + 3\frac{\sigma}{\sqrt{n}} = \bar{\bar{X}} + 3\frac{\bar{S}}{c_4\sqrt{n}} = \bar{\bar{X}} + A_3\bar{S}$$

$$CL_{\bar{X}} = \mu_{\overline{X_i}} = \bar{\bar{X}}$$

$$LCL_{\bar{X}} = \mu_{\overline{X_i}} - 3\sigma_{\overline{X_i}} = \bar{\bar{X}} - 3\frac{\sigma}{\sqrt{n}} = \bar{\bar{X}} - 3\frac{\bar{S}}{c_4\sqrt{n}} = \bar{\bar{X}} - A_3\bar{S}$$

όπου B_3, B_4, A_3 είναι συντελεστές για τα διάφορα μεγέθη δειγμάτων.

Όρια ελέγχου για το $\bar{X}$-διάγραμμα:

$$UCL = \bar{\bar{X}} + A_3\bar{S}$$

$$Center\ Line = \bar{\bar{X}}$$

$$LCL = \bar{\bar{X}} - A_3\bar{S}$$

Όρια ελέγχου για το S-διάγραμμα:

$$UCL = B_4\bar{S}$$

$$Center\ Line = \bar{S}$$

$$LCL = B_3\bar{S}$$

5.1.4 Διάγραμμα συνεχούς μέσου-εύρους

Σε μερικές διεργασίες τα ποιοτικά χαρακτηριστικά παρέχουν μόνο μία τιμή για ανάλυση. Σε αυτές τις περιπτώσεις χρησιμοποιούμε ένα διάγραμμα συνεχούς μέσου-εύρους. Σε ένα τέτοιο διάγραμμα δημιουργείται ένα ψευδές δείγμα από την νέα και από τις πιο πρόσφατες τιμές. Ένα τυπικό μέγεθος δείγματος είναι 3, όπου το δείγμα αποτελείται από την νέα και τις δύο προηγούμενες τιμές του ποιοτικού χαρακτηριστικού. Στην συνέχεια ισχύουν οι συνήθεις οδηγίες μιας μελέτης Στατιστικού Ελέγχου Διεργασιών. Η συνέπεια της δημιουργίας των ψευδών δειγμάτων είναι ότι οι προκύπτουσες στατιστικές συναρτήσεις του δείγματος δεν είναι ανεξάρτητες η μία από την άλλη. Ο καθοριστικός παράγοντας του μεγέθους του δείγματος είναι το χρονικό διάστημα μεταξύ των ενδείξεων του δείγματος, το οποίο μπορεί να είναι τόσο μεγάλο όσο επιτρέπουν οι σχετικές απαιτήσεις.

Αν απαιτείται μεγάλη ευαισθησία ή αν υπάρχει μεγάλη καθυστέρηση μεταξύ των ενδείξεων, ίσως χρειαστεί να σημειωθούν μεμονωμένες τιμές στο διάγραμμα μέσου, παρόλο που ίσως χρειαστεί να χρησιμοποιηθούν στατιστικές συναρτήσεις δείγματος για να ορίσουμε τα όρια ελέγχου. Όσο μεγαλύτερο είναι το μέγεθος του δείγματος τόσο πιο ομαλό είναι το αποτέλεσμα στο διάγραμμα.

5.1.5 Διαγράμματα ελέγχου για μεμονωμένες παρατηρήσεις

Στις περιπτώσεις που το μέγεθος του δείγματος είναι ίσο με 1, χρησιμοποιούνται τα διαγράμματα μεμονωμένων ή ατομικών παρατηρήσεων (individual observations). Αν το χαρακτηριστικό X των προϊόντων που παράγονται ακολουθεί την κανονική κατανομή $N(\mu, \sigma^2)$ με μ, σ γνωστά τότε τα όρια ελέγχου για την μέση τιμή του χαρακτηριστικού X, για n=1 είναι:

$$LCL_{\bar{X}} = \bar{X} - 3\frac{\sigma}{\sqrt{n}} = \bar{X} - 3\sigma$$
$$CL_{\bar{X}} = \bar{X}$$
$$UCL_{\bar{X}} = \bar{X} + 3\frac{\sigma}{\sqrt{n}} = \bar{X} + 3\sigma$$

Οι παρατηρήσεις που απεικονίζονται στο διάγραμμα είναι οι μεμονωμένες.

Επειδή δεν μπορεί να χρησιμοποιηθεί το R διάγραμμα ελέγχου για n=1, χρησιμοποιείται το κινούμενο εύρος (moving range-MR) των μεμονωμένων παρατηρήσεων που δίνεται από την σχέση:

$$MR_i = |X_i - X_{i-1}|, i \geq 2$$

και ισχύει:

$$\mu_{MR_i} = E(MR_i) = \sigma d_2$$
$$\sigma_{MR_i} = \sqrt{Var(MR_i)} = \sigma d_3$$

όπου οι σταθερές d_2, d_3 υπολογίζονται για n=2.

Το μοντέλο των 3-σ ορίων θα είναι:

$$UCL_{MR} = \mu_{MR_i} + 3\sigma_{MR_i} = (d_2 + 3d_3)\sigma = D_2\sigma$$
$$CL_{MR} = \mu_{MR_i} = d_2\sigma$$
$$LCL_{MR} = \mu_{MR_i} - 3\sigma_{MR_i} = (d_2 - 3d_3)\sigma = D_1\sigma$$

όπου οι σταθερές d_2, d_3 υπολογίζονται για n=2.

Στην περίπτωση που τα μ, σ είναι άγνωστα, πρέπει να εκτιμηθούν. Έστω $\boldsymbol{X} = (X_1, X_2, \dots, X_m)$ ένα τυχαίο δείγμα μεγέθους m από το χαρακτηριστικό $X \sim N(\mu, \sigma^2)$. Τότε,

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_m}{m}$$

με

$$E(\bar{X}) = \mu$$
$$Var(\bar{X}) = \frac{\sigma^2}{m}$$

Για την μεταβλητή $\bar{X}$ ισχύει:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{m})$$

και

$$\hat{\mu} = \bar{X}.$$

Θέτοντας

$$\overline{MR} = \frac{MR_1 + MR_2 + \cdots + MR_{m-1}}{m-1}$$

προκύπτει

$$E(\overline{MR}) = \frac{1}{m-1} E(\sum_{i=1}^{m-1} MR_i) = \sigma d_2$$

Επομένως,

$$\hat{\sigma} = \frac{\overline{MR}}{d_2}$$

Τα όρια ελέγχου είναι:

$$LCL_X = \bar{X} - 3\frac{\overline{MR}}{d_2}$$
$$CL_X = \bar{X}$$
$$UCL_X = \bar{X} + 3\frac{\overline{MR}}{d_2}$$

$$LCL_{MR} = \left(1 - 3\frac{d_3}{d_2}\right)\overline{MR} = D_3\overline{MR}$$
$$CL_{MR} = \overline{MR}$$
$$UCL_{MR} = \left(1 + 3\frac{d_3}{d_2}\right)\overline{MR} = D_4\overline{MR}$$

όπου οι σταθερές D_3, D_4 υπολογίζονται για n=2.

Η αποτελεσματικότερη εκτίμηση της τυπικής απόκλισης σ προκύπτει από την σχέση $\hat{\sigma}$ =S/c_4, όπου η σταθερά c_4 υπολογίζεται για $n = m$ και $S = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2}$.

5.1.6 ΣΕΔ για μεγάλα δείγματα

Στις περιπτώσεις που η συλλογή δεδομένων είναι εύκολη και πραγματοποιείται σε μικρό χρονικό διάστημα και σε μεγάλες ποσότητες χρησιμοποιείται ο προσεγγιστικός τύπος:

$$\bar{\bar{X}} \pm 3\sigma_{\bar{X}} = \bar{\bar{X}} \pm 3\frac{\hat{\sigma}}{\sqrt{n}}.$$

Όταν ο αριθμός των δειγμάτων m είναι μεγάλος και υπάρχουν διαφορετικά μεγέθη δειγμάτων $n_1, \dots, n_m$ χρησιμοποιείται ο προσεγγιστικός τύπος:

$$\bar{\bar{X}} \pm 3\sigma_{\bar{X}} = \bar{\bar{X}} \pm 3\frac{S_p}{\sqrt{2\bar{n}}} ,$$

όπου $\bar{n}$ είναι ο μέσος όρος των μεγεθών των δειγμάτων $\bar{n} = \frac{n_1+\cdots+n_m}{m}$, $\bar{\bar{X}}$ ο γενικός μέσος όρος των μέσων των δειγμάτων και S_p μια εκτίμηση της συνδυασμένης τυπικής απόκλισης της διεργασίας που μπορεί να υπολογιστεί από την σχέση:

$$S_p = \sqrt{\frac{(n_1-1)s_1^2 + \cdots + (n_m-1)s_m^2}{(n_1-1)+\cdots+(n_m-1)}}.$$

Θεωρώντας κανονική κατανομή, τα όρια ελέγχου της μεταβλητότητας της διαδικασίας προσεγγίζονται από τον τύπο:

$$S_p \cdot \frac{\sqrt{2\bar{n}-3} \pm 3}{\sqrt{2\bar{n}-2}}.$$

5.2 Διαγράμματα ελέγχου ιδιοτήτων

5.2.1 Εισαγωγή

Τα διαγράμματα ελέγχου ιδιοτήτων (attribute control charts) χρησιμοποιούνται όταν τα δεδομένα αφορούν ποσότητες με δύο πιθανά αποτελέσματα της μορφής ναι / όχι. Το χαρακτηριστικό της ποιότητας είναι είτε οι ελαττωματικές μονάδες, είτε τα ελαττώματα μιας μονάδας. Το μέγεθος του δείγματος μπορεί να είναι σταθερό ή να μεταβάλλεται από δείγμα σε δείγμα. Ένα προϊόν ταξινομείται σαν ελαττωματικό ή μη συμμορφούμενο (defective/ nonconforming) αν τουλάχιστον ένα ποιοτικό χαρακτηριστικό του έχει τιμή εκτός των ορίων προδιαγραφών, δηλαδή το προϊόν παρουσιάζει ελάττωμα ή ατέλεια (defect/ nonconformity). Μία μονάδα μπορεί να παρουσιάσει πολλά ελαττώματα, όχι απαραίτητα του ίδιου τύπου, για να ταξινομηθεί σαν ελαττωματική. Ο όρος μονάδα επιθεώρησης (inspection unit) αναφέρεται στο ίδιο το προϊόν, ή σε τμήμα του προϊόντος ή σε ένα σύνολο προϊόντων.

5.2.2 Διαγράμματα ελέγχου p και np

Το ποιοτικό χαρακτηριστικό που μας ενδιαφέρει στα p και np διαγράμματα ελέγχου είναι αντίστοιχα το ποσοστό και ο αριθμός των ελαττωματικών προϊόντων. Ποσοστό των ελαττωματικών προϊόντων καλείται το πηλίκο του αριθμού των ελαττωματικών προϊόντων προς τον συνολικό αριθμό των παραγόμενων προϊόντων. Για το σχεδιασμό των διαγραμμάτων αρχικά επιλέγονται m προκαταρτικά δείγματα ισομεγέθη ή μη. Τα δείγματα από την παραγωγή δεν έχουν όλα το ίδιο μέγεθος. Αρχικά γίνεται η υπόθεση ότι το ποσοστό των ελαττωματικών προϊόντων μιας παραγωγικής διαδικασίας είναι γνωστό και ίσο με p και ότι από την παραγωγή επιλέγουμε m ανεξάρτητα τυχαία

δείγματα με μεγέθη $n_1, \dots, n_m$ αντίστοιχα. Έστω $X_{ij}, i \geq 1, 1 \leq j \leq n_i$ η τυχαία μεταβλητή με τιμές 1 και 0 ανάλογα με το αν το j προϊόν του i δείγματος είναι ελαττωματικό ή όχι. Για την τυχαία μεταβλητή X_{ij} ισχύει: $X_{ij} \sim B(1, p)$. Για την μεταβλητή

$$X_i = X_{i1} + X_{i2} + \cdots + X_{in_i}, \qquad 1 \leq i \leq m$$

που δηλώνει τον αριθμό των ελαττωματικών προϊόντων στο i δείγμα ισχύει: $X_i \sim B(n_i, p)$, με μέση τιμή $\mu_{X_i} = n_i p$ και διασπορά $\sigma^2_{X_i} = n_i p(1-p)$. Τότε για την τυχαία μεταβλητή

$$W_i = p_i = \frac{X_i}{n_i}, \qquad 1 \leq i \leq m$$

που δηλώνει το ποσοστό των ελαττωματικών προϊόντων στο i δείγμα ισχύει:

$$\mu_{W_i} = E(W_i) = \frac{1}{n_i} E(X_i) = p \quad \kappa\alpha\iota$$

$$\sigma^2_{W_i} = Var(W_i) = \frac{1}{{n_i}^2} Var(X_i) = \frac{p(1-p)}{n_i}, i \geq 1.$$

Το διάγραμμα ελέγχου που θα χρησιμοποιείται για την παρακολούθηση του ποσοστού των ελαττωματικών προϊόντων θα είναι η απεικόνιση της στατιστικής συνάρτησης $W_i = p_i = \frac{X_i}{n_i}$. Τα όρια ελέγχου και η κεντρική γραμμή του p διαγράμματος ελέγχου, για την περίπτωση που το ποσοστό p των ελαττωματικών προϊόντων της διεργασίας είναι γνωστό, είναι:

$$UCL_p = \mu_{W_i} + 3\sigma_{W_i} = p + 3\sqrt{\frac{p(1-p)}{\bar{n}}}$$

$$CL_p = \mu_{W_i} = p$$

$$LCL_p = \mu_{W_i} - 3\sigma_{W_i} = p - 3\sqrt{\frac{p(1-p)}{\bar{n}}}$$

όπου $\bar{n} = \frac{n_1 + \cdots + n_m}{m}$ ο μέσος όρος όλων των μεγεθών των m δειγμάτων (για ισομεγέθη δείγματα, $\bar{n} = n$). Στην περίπτωση που τα δείγματα είναι ισομεγέθη, η κατασκευή ενός διαγράμματος ελέγχου για την παρακολούθηση του αριθμού των ελαττωματικών προϊόντων επιταχύνεται με την βοήθεια της τυχαίας μεταβλητής X_i. Στο διάγραμμα ελέγχου θα απεικονίζεται η στατιστική συνάρτηση X_i για την οποία $\mu_{X_i} = np$, $\sigma^2_{X_i} = np(1-p)$, $1 \leq i \leq m$. Τα όρια και η κεντρική γραμμή για το np διάγραμμα ελέγχου με γνωστό p, είναι:

$$UCL_{np} = \mu_{X_i} + 3\sigma_{X_i} = np + 3\sqrt{np(1-p)}$$

$$CL_{np} = \mu_{X_i} = np$$

$$LCL_{np} = \mu_{X_i} - 3\sigma_{X_i} = np - 3\sqrt{np(1-p)}$$

Όταν το ποσοστό p των ελαττωματικών προϊόντων της διεργασίας είναι άγνωστο πρέπει να εκτιμηθεί. Έστω m ανεξάρτητα προκαταρκτικά δείγματα με μεγέθη $n_1, \dots, n_m$ αντίστοιχα, έστω τα $X_i = X_{i1}, X_{i2}, \dots, X_{in_i}$, $1 \le i \le m, X_i \sim B(n_i, p)$, με μέση τιμή $\mu_{X_i} = n_i p$ και διασπορά $\sigma^2_{X_i} = n_i p(1-p)$.
Θέτουμε

$$W_i = p_i = \frac{X_i}{n_i}, \qquad 1 \le i \le m$$

και προκύπτει

$$\bar{p} = \frac{X_1 + X_2 + \cdots + X_m}{n_1 + \cdots + n_m}$$

με

$$E(\bar{p}) = p$$

Δηλαδή $\hat{p} = p$.
Τα όρια ελέγχου p διαγράμματος είναι τα εξής:

$$UCL_p = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{\bar{n}}}$$

$$CL_p = \bar{p}$$

$$LCL_p = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{\bar{n}}}$$

Τα όρια ελέγχου np διαγράμματος είναι τα εξής:

$$UCL_{np} = n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})}$$

$$CL_{np} = n\bar{p}$$

$$LCL_{np} = n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})}$$

Αν το LCL είναι αρνητικό τότε η τιμή θεωρείται μηδέν.

5.2.3 Διάγραμμα ελέγχου c

Το ποιοτικό χαρακτηριστικό με το οποίο ασχολείται το διάγραμμα c είναι ο συνολικός αριθμός ελαττωμάτων σε μία μονάδα επιθεώρησης. Η βασική υπόθεση είναι ότι ο συνολικός αριθμός των ελαττωμάτων μιας μονάδας ακολουθεί την κατανομή *Poisson* και σύμφωνα με αυτήν την υπόθεση, η πιθανότητα εμφάνισης ελαττώματος σε οποιοδήποτε σημείο μιας μονάδας θα πρέπει να είναι πολύ μικρή. Έτσι ο αριθμός X των ελαττωμάτων που εμφανίζονται σε μια μονάδα επιθεώρησης ακολουθεί την κατανομή Poisson με παράμετρο c. Δηλαδή $X \sim P(c)$ και $P(X = x) = e^{-c}\frac{c^x}{x!}, x = 0,1, \dots$ με $\mu_X = \sigma^2_X = c$. Στο διάγραμμα ελέγχου για την παρακολούθηση του αριθμού των ελαττωμάτων θα απεικονίζεται η στατιστική συνάρτηση $X_i, i \ge 1$ όπου X_i δηλώνει τον αριθμό των ελαττωμάτων στην i μονάδα επιθεώρησης. Τα όρια ελέγχου και η κεντρική γραμμή όταν το c είναι γνωστό είναι:

$$UCL_c = c + 3\sqrt{c}$$

$$CL_c = c$$

$$LCL_c = c - 3\sqrt{c}$$

Όταν η παράμετρος c της κατανομής *Poisson* είναι άγνωστη, πρέπει να εκτιμηθεί. Έστω X_i ο αριθμός των ελαττωμάτων στην i μονάδα επιθεώρησης, $1 \leq i \leq m$. Θέτουμε

$$\bar{c} = \frac{X_1 + X_2 + \cdots + X_m}{m}\ , 1 \leq i \leq m$$

και προκύπτει

$$E(\bar{c}) = c$$

Δηλαδή $\hat{c} = \bar{c}$.
Τα όρια ελέγχου του c διαγράμματος είναι τα εξής:

$$UCL_c = \bar{c} + 3\sqrt{\bar{c}}$$
$$CL_c = \bar{c}$$
$$LCL_c = \bar{c} - 3\sqrt{\bar{c}}$$

Αν το LCL είναι αρνητικό τότε η τιμή θεωρείται μηδέν.

Το σφάλμα τύπου I σε ένα c διάγραμμα ελέγχου είναι

$$\alpha = P(X < LCL_c \text{ ή } X > UCL_c | X \sim P(c))$$

όπου c είναι η εντός ελέγχου τιμή της παραμέτρου της κατανομής *Poisson*.
Το σφάλμα τύπου II σε ένα c διάγραμμα ελέγχου είναι

$$\beta = P(LCL_c \leq X \leq UCL_c | X \sim P(c))$$

όπου c είναι μία εκτός ελέγχου τιμή της παραμέτρου της κατανομής *Poisson*.

5.2.4 Διάγραμμα ελέγχου u

Το ποιοτικό χαρακτηριστικό με το οποίο ασχολείται το διάγραμμα u είναι η αναλογία των ελαττωμάτων ανά δείγμα, ή αλλιώς ο μέσος αριθμός ελαττωμάτων ανά μονάδα επιθεώρησης σε κάθε δείγμα. Η βασική υπόθεση είναι ότι ο αριθμός X των ελαττωμάτων μιας μονάδας ακολουθεί την κατανομή *Poisson* με παράμετρο c. Η βασική διαφορά με το c διάγραμμα ελέγχου είναι ότι στα u διαγράμματα ελέγχου μπορούμε να έχουμε δείγματα μεγέθους μεγαλύτερου της μιας μονάδας επιθεώρησης. Έστω ότι από την παραγωγή επιλέγουμε m ανεξάρτητα τυχαία δείγματα με μεγέθη $n_1, n_2, \ldots, n_m$ μονάδες επιθεώρησης αντίστοιχα.
Συμβολίζουμε με $X_{ij}, i \geq 1,\ 1 \leq j \leq n_i$ την τυχαία μεταβλητή που δηλώνει τον αριθμό των ελαττωμάτων της j μονάδας επιθεώρησης στο i δείγμα. Για την τυχαία μεταβλητή X_{ij} έχουμε ότι $X_{ij} \sim P(c)$ με $\mu_{X_{ij}} = \sigma^2_{X_{ij}} = c$ ενώ για την

$$X_i = X_{i1} + X_{i2} + \cdots + X_{in_i}\ , \qquad 1 \leq i \leq m$$

που δηλώνει τον αριθμό των ελαττωμάτων στο i δείγμα έχουμε ότι $X_i \sim P(n_i c)$ με $\mu_{X_i} = \sigma^2_{X_i} = n_i c$. Τότε για την τυχαία μεταβλητή

$$u_i = \frac{X_i}{n_i}, 1 \leq i \leq m$$

που δηλώνει την αναλογία των ελαττωμάτων στο i δείγμα ισχύει

$$\mu_{u_i} = c$$

$\sigma^2_{u_i} = \frac{c}{n_i}, i \geq 1.$

Συνεπώς μπορούμε να αναπτύξουμε ένα διάγραμμα ελέγχου για την παρακολούθηση του μέσου αριθμού των ελαττωμάτων ανά μονάδα επιθεώρησης σε κάθε δείγμα, στο οποίο θα απεικονίζεται η στατιστική συνάρτηση $u_i = {X_i}/{n_i}$. Όταν η παράμετρος c της κατανομής είναι γνωστή, τα όρια και η κεντρική γραμμή του διαγράμματος ελέγχου είναι τα εξής:

$$UCL_u = c + 3\sqrt{\frac{c}{\bar{n}}}$$
$$CL_u = c$$
$$LCL_u = c - 3\sqrt{\frac{c}{\bar{n}}}$$

όπου $\bar{n} = \frac{n_1 + \cdots + n_m}{m}$ ο μέσος όρος όλων των μεγεθών των m δειγμάτων (για ισομεγέθη δείγματα, $\bar{n} = n$).

Όταν η παράμετρος c της κατανομής *Poisson* είναι άγνωστη, πρέπει να εκτιμηθεί. Έστω X_i ο αριθμός των ελαττωμάτων στην i μονάδα επιθεώρησης, $1 \leq i \leq m$. Θέτουμε

$$u_i = \frac{X_i}{n_i}, 1 \leq i \leq m$$

$$\bar{u} = \frac{X_1 + X_2 + \cdots + X_m}{n_1 + \cdots + n_m}, 1 \leq i \leq m$$

και προκύπτει

$$E(\bar{u}) = c$$

Δηλαδή $\hat{c} = \bar{u}$.

Τα όρια ελέγχου του u διαγράμματος είναι τα εξής:

$$UCL_u = \bar{u} + 3\sqrt{\frac{\bar{u}}{\bar{n}}}$$
$$CL_u = \bar{u}$$
$$LCL_u = \bar{u} - 3\sqrt{\frac{\bar{u}}{\bar{n}}}$$

Αν το LCL είναι αρνητικό τότε η τιμή θεωρείται μηδέν.

Το σφάλμα τύπου I σε ένα c διάγραμμα ελέγχου είναι

$$\alpha = P(\frac{X_i}{n_i} < LCL_u \text{ ή } \frac{X_i}{n_i} > UCL_u | X_i \sim P(n_i c))$$

όπου c είναι η εντός ελέγχου τιμή της παραμέτρου της κατανομής *Poisson*.
Το σφάλμα τύπου II σε ένα c διάγραμμα ελέγχου είναι

$$\beta = P(LCL_u \leq \frac{X_i}{n_i} \leq UCL_u | X_i \sim P(n_i c))$$

όπου c είναι μία εκτός ελέγχου τιμή της παραμέτρου της κατανομής *Poisson*.

5.3 Αθροιστικά διαγράμματα ελέγχου

5.3.1 Εισαγωγή

Τα αθροιστικά διαγράμματα ελέγχου εισήχθησαν το 1954 από τον Βρετανό χημικό Page (Page E., 1961) και είναι γνωστά ως διαγράμματα Cusum (Cumulative Sum). Το βασικό πλεονέκτημα είναι ότι είναι ευαίσθητα στον εντοπισμό μικρών εκτροπών και χρησιμοποιούνται για αυτό τον λόγο στην ανίχνευση μικρών συστηματικών σφαλμάτων. Το χαρακτηριστικό X που μας ενδιαφέρει ακολουθεί την κανονική κατανομή $N(\mu_0, \sigma^2)$. Επιλέγοντας m ανεξάρτητα τυχαία δείγματα $X_i = (X_{i1}, \dots, X_{in}), 1 \leq i \leq m$ μεγέθους $n \geq 1$ το καθένα, ο δειγματικός μέσος

$$\bar{X}_i = \frac{X_{i1} + \cdots + X_{in}}{n}$$

ακολουθεί κανονική κατανομή με μέση τιμή $\mu_{\bar{X}_i} = \mu_0$ και διασπορά $\sigma^2_{\bar{X}_i} = {\sigma^2}/{n}$.

Τα διαγράμματα Cusum είναι αποδοτικά στις περιπτώσεις μεμονωμένων παρατηρήσεων και ενσωματώνουν άμεσα όλες τις πληροφορίες της ακολουθίας των παρατηρήσεων, γιατί παριστάνουν τα συσσωρευμένα αθροίσματα C_i των διαφόρων δειγματικών τιμών από την τιμή στόχο μ_0.

5.3.2 Διάγραμμα Tabular Cusum

Στο διάγραμμα Tabular Cusum είναι απαραίτητος ο υπολογισμός δύο συσσωρευτικών αθροισμάτων για κάθε τιμή ελέγχου. Οι θετικές αποκλίσεις από τον στόχο συναθροίζονται με το άνω συσσωρευμένο άθροισμα C_i^+(one-sided upper cusum), ενώ οι αρνητικές αποκλίσεις από τον στόχο συναθροίζονται με το κάτω συσσωρευμένο άθροισμα C_i^-(one-sided lower cusum), σύμφωνα με τις παρακάτω σχέσεις:

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

με $1 \leq i \leq m$ και αρχικές τιμές $C_0^+ = C_0^- = 0$.

Τα αθροίσματα C_i^+, C_i^- υπολογίζονται από τις διαφορές των τιμών X_i από την μέση τιμή μ_0, εφόσον αυτές είναι μεγαλύτερες από την τιμή αναφοράς K. Κάθε φορά που οι διαφορές γίνονται αρνητικές, το άθροισμα (C_i^+ ή C_i^-) μηδενίζεται για να ξαναρχίσει να αυξάνεται όταν οι διαφορές γίνουν ξανά μεγαλύτερες του μηδενός. Στο διάγραμμα Tabular Cusum τα αθροίσματα C_i^+, C_i^- σχεδιάζονται ως δύο διαφορετικές στήλες πάνω και κάτω από την μέση τιμή. Η τιμή αναφοράς K δεν σχεδιάζεται στο διάγραμμα Tabular Cusum. Το όριο ελέγχου που σχεδιάζεται στο διάγραμμα είναι το διάστημα απόφασης H (decision interval). Το ανώτερο H^+ και το κατώτερο H^- σχεδιάζονται με δύο ευθείες γραμμές παράλληλες προς το μέσο μ_0. Η τιμή του διαστήματος απόφασης υποδεικνύει τα ανώτατα επιτρεπτά όρια των αθροισμάτων C_i^+, C_i^- και επιλέγεται να είναι $H = 5\sigma$.

Θέτουμε $H = h\sigma$, $K = k\sigma = \frac{\delta}{2}\sigma$, $\delta = \frac{|\mu_1 - \mu_0|}{\sigma}$, όπου k το μέγεθος της μετατόπισης που θέλουμε να ανιχνευτεί, μ_0 η τιμή στόχος (μέση τιμή), μ_1 η εκτός ελέγχου τιμή του μέσου (η ανώτατη επιτρεπτή τιμή των δειγμάτων ελέγχου). Η επιλογή της παραμέτρου k εξαρτάται από το μέγεθος της μετατόπισης που θέλουμε να ανιχνευτεί.

5.3.3 Τυποποιημένο διάγραμμα Cusum

Σε μερικές περιπτώσεις είναι προτιμότερο να τυποποιείται η μεταβλητή X_i πριν τους υπολογισμούς των συσσωρευμένων αθροισμάτων. Ορίζεται η μεταβλητή

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

η οποία αποτελεί την τυποποιημένη τιμή της X_i.

Τα άνω και κάτω συσσωρευτικά αθροίσματα μετασχηματίζονται ως εξής:

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

με $1 \le i \le m$ και αρχικές τιμές $C_0^+ = C_0^- = 0$.

Πλεονεκτήματα των τυποποιημένων διαγραμμάτων Cusum είναι τα παρακάτω:

i. Δυνατότητα ύπαρξης πολλών διαγραμμάτων Cusum με τις ίδιες τιμές των k και h καθώς οι επιλογές των παραμέτρων δεν αποτελούν ακολουθία εξαρτημένων τιμών. Οι παράμετροι δηλαδή δεν εξαρτώνται από την τυπική απόκλιση σ της κάθε διεργασίας γιατί στην τυποποιημένη κανονική κατανομή $\sigma = 1$.
ii. Με την χρήση της τυποποιημένης μεταβλητής Y_i δημιουργούνται εύκολα τα διαγράμματα για τον έλεγχο της μεταβλητότητας μιας διαδικασίας.

5.3.4 Διάγραμμα Scale Cusum

Τα διαγράμματα Scale Cusum χρησιμοποιούνται για την παρακολούθηση της μεταβλητότητας που ενέχει μια διαδικασία. Έστω όπως παραπάνω, η κανονική μεταβλητή X_i, με μέση τιμή ή τιμή στόχο μ_0 και τυπική απόκλιση σ. Η τυποποιημένη τιμή της X_i είναι η $Y_i = {X_i - \mu_0}/{\sigma}$. Με την νέα μεταβλητή v_i

(Hawkins, 1991),

$$v_i = \frac{\sqrt{|Y_i|} - 0.822}{0.349}, \qquad 1 \le i \le m$$

οι ποσότητες v_i είναι ευαίσθητες σε αλλαγές της διασποράς μιας διαδικασίας και ακολουθούν την τυποποιημένη κανονική κατανομή $N(0,1)$. Τα δύο μονόπλευρα Scale Cusums είναι τα εξής:

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

με $1 \le i \le m$ και αρχικές τιμές $S_0^+ = S_0^- = 0$.

Αν η τυπική απόκλιση της διεργασίας αυξάνεται, τότε θα αυξάνονται και οι τιμές των αθροισμάτων S_i^+ ξεπερνώντας κάποια στιγμή το διάστημα απόφασης H, ενώ αν η τυπική απόκλιση μειώνεται θα μειώνονται και οι τιμές των S_i^- μέχρι τελικά να ξεπεράσουν την τιμή H. Οι τιμές των παραμέτρων k και h επιλέγονται όπως στα διαγράμματα του μέσου.
Αν υπάρξει ένδειξη εκτός ελέγχου στο διάγραμμα Scale Cusum τότε υπάρχει υποψία αλλαγής στην διασπορά της διαδικασίας, ενώ αν υπάρξει ένδειξη εκτός ελέγχου και στα δύο διαγράμματα (διαγράμματα μέσου και μεταβλητότητας) τότε υπάρχει υποψία μετατόπισης στο μέσο.

5.4 Διαγράμματα ελέγχου με κινητούς μέσους και εκθετικά βάρη (Exponentially Weighted Moving Average Control Charts – EWMA)

5.4.1 Εισαγωγή

Το διάγραμμα EWMA είναι ένα εναλλακτικό διάγραμμα των διαγραμμάτων Shewart και παρουσιάστηκε το 1959 από τον Roberts (Roberts S.W., 1959). Η χρήση του συνίσταται κυρίως στην περίπτωση που θέλουμε να εντοπίσουμε μικρές μεταβολές στο μέσο μιας διαδικασίας και χρησιμοποιείται επίσης για μεμονωμένες παρατηρήσεις. Οι τιμές προς έλεγχο τοποθετούνται πάνω στο διάγραμμα ως z_i, τα οποία υπολογίζονται βάσει του τύπου

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

όπου x_i είναι οι παρατηρήσεις, λ είναι μια παράμετρος που ονομάζεται συντελεστής βαρύτητας (weighting factor), με $\lambda \in (0,1]$. Η παράμετρος αυτή καθορίζει το βαθμό κατά τον οποίο παλαιότερα δεδομένα εισάγονται στον υπολογισμό του EWMA στατιστικού. Όσο πιο κοντά στην μονάδα είναι η τιμή του λ τόσο μικρότερο βάρος δίδεται στα προγενέστερα δεδομένα. Η πρώτη τιμή $z_0 = \mu_0$ είναι η μέση τιμή–τιμή στόχος (target value). Ως αρχική τιμή μπορούμε να χρησιμοποιήσουμε το μέσο από τα προηγούμενα δεδομένα (historical data) ή το $\bar{\bar{x}}$ (δειγματικός μέσος). Ο EWMA είναι τελικά ένας σταθμισμένος μέσος όλων των προηγούμενων παρατηρήσεων ο οποίος περιγράφεται ως εξής:

$$z_i = \lambda \sum_{j=0}^{i-1} (1-\lambda)^j x_{i-j} + (1-l)^i z_0$$

Θεωρώντας ότι τα x_i είναι ανεξάρτητες τυχαίες μεταβλητές με διασπορά σ^2, η διασπορά των z_i είναι $\sigma_{z_i}^2 = \sigma^2 \left(\frac{\lambda}{2-\lambda}\right)\left(1-(1-\lambda)^{2i}\right)$. Το διάγραμμα EWMA μπορεί να κατασκευαστεί κάνοντας το γράφημα των z_i ως προς τον δειγματικό αριθμό i. Τα όρια ελέγχου (control limits) είναι:

$$UCL = \mu_0 + L\sigma\sqrt{\left(\frac{\lambda}{2-\lambda}\right)(1-(1-\lambda)^{2i})}$$

$$CL = \mu_0$$

$$LCL = \mu_0 - L\sigma\sqrt{\left(\frac{\lambda}{2-\lambda}\right)(1-(1-\lambda)^{2i})}$$

όπου L το εύρος των ορίων και μ_0 η τιμή στόχος.
Επειδή $1-(1-\lambda)^{2i} \to 1$ καθώς αυξάνεται το i, τα όρια μετατρέπονται:

$$UCL = \mu_0 + L\sigma\sqrt{\left(\frac{\lambda}{2-\lambda}\right)}$$

$$CL = \mu_0$$

$$LCL = \mu_0 - L\sigma\sqrt{\left(\frac{\lambda}{2-\lambda}\right)}$$

Τα άνω και κάτω όρια ελέγχου αποκτούν μια σταθερή τιμή και στο διάγραμμα ελέγχου απεικονίζονται ως δύο ευθείες γραμμές παράλληλες μεταξύ τους. Τα παραπάνω ισχύουν και στην περίπτωση που οι παρατηρήσεις είναι μεμονωμένες.

5.4.2 Σχεδιασμός

Το ARL (Average Run Length –Μέσο μήκος ροής) ενός διαγράμματος ελέγχου ορίζεται ως ο αναμενόμενος αριθμός σημείων που πρέπει να απεικονιστούν στο

διάγραμμα, μέχρις ότου να υπάρξει ένδειξη εκτός ελέγχου. Χωρίζεται σε εντός ελέγχου (in-control) ARL_0 και εκτός ελέγχου (out-of-control) ARL_1. Το ARL_0 είναι προτιμότερο να είναι μεγάλο ώστε σε μια εντός ελέγχου διαδικασία να χρειαστεί μεγάλος αριθμός δειγμάτων μέχρι την εσφαλμένη ένδειξη ότι η διαδικασία είναι εκτός ελέγχου. Αντιθέτως, το ARL_1 είναι προτιμότερο να είναι μικρό, γιατί σε μια εκτός ελέγχου διαδικασία πρέπει από όσο γίνεται μικρό αριθμό δειγμάτων να προκύψει ένδειξη ότι η διαδικασία είναι εκτός ελέγχου. Το πρόβλημα στο EWMA έγκειται στην σωστή επιλογή των παραμέτρων λ και L, ώστε να έχει η ARL απόδοση την επιθυμητή τιμή. Όσον αφορά το λ, συνίσταται η χρήση μικρών τιμών του για την ανίχνευση μικρών αλλαγών στην διαδικασία, ενώ για το L χρησιμοποιείται συνήθως η τιμή L=3. Μεγαλώνοντας το πλάτος των ορίων, μειώνεται το σφάλμα τύπου I, δηλαδή να προκύψει σημείο εκτός ελέγχου ενώ στην πραγματικότητα είναι εντός. Αυξάνεται όμως η πιθανότητα σφάλματος τύπου II, δηλαδή σημεία που φαίνονται ότι είναι εντός στην πραγματικότητα να είναι εκτός. Έχουν γίνει αρκετές έρευνες σχετικά με την ARL συμπεριφορά του διαγράμματος σε σχέση με τις τιμές λ και L. Κατά τους Lucas και Saccucci (1990) (Lucas J.M. and M.S. Saccucci), το βέλτιστο διάγραμμα EWMA με κριτήριο την τιμή του ARL, επιτυγχάνεται όταν:

- Έχουμε καθορίσει το επιθυμητό εντός ελέγχου ARL.
- Έχουμε καθορίσει το μέγεθος της αλλαγής στο μέσο της διαδικασίας που θέλουμε να εντοπίσουμε.
- Από τα διάφορα ζεύγη λ και L, επιλέγουμε εκείνο που δίνει την μικρότερη τιμή ARL για την συγκεκριμένη αλλαγή μου μας ενδιαφέρει.

5.4.3 Μειονεκτήματα

Στην περίπτωση που έχουμε δώσει μικρή τιμή στο λ, παρουσιάζεται το παρακάτω πρόβλημα. Αν η τιμή του EWMA στατιστικού είναι από την μία πλευρά της κεντρικής γραμμής και προκύψει μεταβολή στο μέσο στην αντίθετη κατεύθυνση, αυτό θα οδηγήσει στην ανάγκη χρονικών περιόδων ώστε να αντιδράσει το EWMA στην μεταβολή αυτή για να το εντοπίσει. Αυτό διότι η μικρή τιμή του λ, δεν δίνει μεγάλο βάρος στις νέες παρατηρήσεις. Αυτό το φαινόμενο είναι γνωστό ως φαινόμενο αδράνειας (inertia effect). Επίσης, το EWMA δεν αντιδρά γρήγορα σε μεγάλες αλλαγές στο μέσο, συγκριτικά με το Shewhart διάγραμμα. Ένας τρόπος να αυξήσουμε την ευαισθησία του διαγράμματος στην ανίχνευσή τους, χωρίς να θυσιάσουμε την ικανότητά του στην ανίχνευση μικρών αλλαγών, είναι να γίνει ένας συνδυασμός ενός Shewhart και ενός διαγράμματος EWMA. Το συνδυασμένο διάγραμμα που προκύπτει, θα είναι εξίσου αποτελεσματικό τόσο στις μικρές όσο και στις μεγάλες μεταβολές στο μέσο, ιδίως στην περίπτωση που χρησιμοποιηθούν ελαφρώς μεγαλύτερα όρια στο Shewhart διάγραμμα.

6 ΚΕΦΑΛΑΙΟ – ΜΗ ΠΑΡΑΜΕΤΡΙΚΑ ΔΙΑΓΡΑΜΜΑΤΑ ΕΛΕΓΧΟΥ

6.1 Στατιστικός έλεγχος διεργασίας-Statistical Process Control (SPC)

Ο στατιστικός έλεγχος διεργασίας (ΣΕΔ) είναι μία από τις πιο διαδεδομένες τεχνικές για τον έλεγχο ποιότητας. Το βασικό αντικείμενο του ΣΕΔ είναι να εντοπίζει γρήγορα την εμφάνιση της αιτίας παραλλαγής, έτσι ώστε η διαδικασία να μπορεί να εξεταστεί και να μπορούν να γίνουν οι απαραίτητες επιδιορθωτικές ενέργειες πριν χειροτερέψει η ποιότητα και παραχθούν ελαττωματικές μονάδες (Stoumbos *et al .,* 2000*)* Βασικό εργαλείο του ΣΕΔ είναι τα διαγράμματα ελέγχου που παρακολουθούν την εκτέλεση της διεργασίας στο χρόνο με στόχο να διατηρήσουν την διαδικασία εντός ελέγχου. Τα προβλήματα διαγραμμάτων ελέγχου μπορούν να χωριστούν σε δύο φάσεις: Φάση I και Φάση II (Woodall 2000 Woodall και Montgomery 1999).

Η ανάλυση της Φάσης I προσπαθεί να απομονώσει τα εντός-ελέγχου δεδομένα από ένα σύνολο αγνώστων ιστορικών δεδομένων και να θέσει τα όρια του ελέγχου για μελλοντική παρακολούθηση (Zhang and Albin, 2007). Η ανάλυση της Φάσης II παρακολουθεί την μέθοδο χρησιμοποιώντας διαγράμματα ελέγχου που αποκομίζονται από το ξεκαθαρισμένο σύνολο εντός-ελέγχου δεδομένων το οποίο προκύπτει από την ανάλυση της Φάσης I. Με έναν απλό σχεδιασμό ενός διαγράμματος ενός συνόλου στατιστικών στοιχείων παρακολούθησης που αντλούνται από τα αρχικά δείγματα, το διάγραμμα ελέγχου μπορεί να καθορίσει αποτελεσματικά αν η διεργασία είναι εντός-ελέγχου ή όχι. Παραδείγματα στατιστικών στοιχείων παρακολούθησης αποτελούν ο δειγματικός μέσος και το δειγματικό εύρος. Ένα επίσης σημαντικό συστατικό των διαγραμμάτων ελέγχου είναι τα όρια ελέγχου, τα οποία συνήθως υπολογίζονται βάσει της κατανομής των στοιχείων παρακολούθησης.

Ο Hotelling (Hotelling, 1947) επέκτεινε το μονομεταβλητό διάγραμμα ελέγχου για να διαχειριστεί πολυμεταβλητά προβλήματα. Το διάγραμμα T^2 είναι ένα πολυμεταβλητό διάγραμμα ελέγχου που μπορεί να παρακολουθεί μία πολυμεταβλητή διεργασία αποτελεσματικά. Τα T^2 διαγράμματα χρησιμοποιούν το T^2 στατιστικό που υπολογίζεται από την παρακάτω εξίσωση:

$$T^2 = (x - \bar{x})^T S^{-1} (x - \bar{x}) \qquad (1)$$

όπου $\bar{x}$ είναι το διάνυσμα του δειγματικού μέσου και S ο πίνακας της δειγματικής διασποράς, τα οποία έχουν υπολογιστεί από τα εντός-ελέγχου δεδομένα της Φάσης I.

Το στατιστικό T^2 μετρά την απόσταση μεταξύ μιας παρατήρησης και του κλιμακωτού μέσου που έχει υπολογιστεί από τα εντός-ελέγχου δεδομένα. Δοθέντος ότι το x ακολουθεί μια πολυμεταβλητή κανονική κατανομή, το T^2 στατιστικό ακολουθεί μια *F* κατανομή (Mason and Young, 2002). Στα διαγράμματα T^2 η 100α%

περιοχή ουράς μιας F κατανομής χρησιμοποιείται σαν όριο ελέγχου, όπου το α είναι το επίπεδο σημαντικότητας το οποίο καθορίζεται από τον πειραματιστή. Τα T^2 διαγράμματα μπορούν να ελέγξουν αποτελεσματικά τα επίπεδα σφάλματος τύπου I και II όταν η κατανομή των δεδομένων της διαδικασίας είναι η πολυμεταβλητή κανονική κατανομή (Lowry and Montgomery, 1995). Η παραδοχή περί της κατανομής που χρησιμοποιείται στα T^2 διαγράμματα περιορίζει την εφαρμοσιμότητά τους σε μη κανονικά δεδομένα, που μπορεί να προκύψουν σε πολλές σύγχρονες βιομηχανίες. (Bakir, 2006; Charaborti *et al.*, 2001; Kim *et al.*, 2004; Qui, 2008)

Καθώς οι περιορισμοί των παρόντων τεχνικών του ΣΕΔ γίνονται φανεροί σε πιο σύνθετες διεργασίες, οι αλγόριθμοι εξόρυξης δεδομένων εξαιτίας των αποδεδειγμένων ικανοτήτων τους να αναλύουν αποτελεσματικά και να ελέγχουν πάρα πολλά δεδομένα, έχουν την δυνατότητα να επιλύουν σημαντικά προβλήματα στον ΣΕΔ. Παρά το γεγονός ότι υπάρχει τεράστια ποικιλία από μελέτες που σχετίζονται με εφαρμογές των τεχνικών εξόρυξης δεδομένων σε διάφορα επιστημονικά πεδία , ελάχιστες προσπάθειες έχουν γίνει για να ενσωματωθούν οι αλγόριθμοι εξόρυξης δεδομένων στο στατιστικό έλεγχο ποιότητας.(Chinnam, 2002; Cook and Chiu, 1998; Hu *et al.*, 2007; Hwang *et al.*, 2005; Smith, 1994).

Οι μέθοδοι ταξινόμησης μίας τάξης έχουν κοινό στόχο με τα διαγράμματα ελέγχου επειδή και οι δύο μέθοδοι υποθέτουν ότι το εντός-ελέγχου σύνολο (το σύνολο στόχος) είναι ο μόνος πληθυσμός που μπορεί να χρησιμοποιηθεί στην μέτρηση του βαθμού της αντικανονικότητας των νέων παρατηρήσεων. Αρκετές μελέτες έχουν λάβει χώρα πρόσφατα με σκοπό την υλοποίηση-ενσωμάτωση των αλγορίθμων ταξινόμησης μίας κλάσης σαν εναλλακτική στα παραδοσιακά διαγράμματα ελέγχου. Οι Sun και Tsung (2003)πρότειναν διαγράμματα που βασίζονται στην απόσταση από τον πυρήνα (K διαγράμματα), τα οποία βασίζονται σε έναν αλγόριθμο περιγραφής διανυσμάτων υποστήριξης δεδομένων (SVDD-Support Vector Data Description). Ο SVDD είναι μία τροποποιημένη εκδοχή των μηχανών διανυσμάτων υποστήριξης (Support Vector Machine-SVM) για την επίλυση προβλημάτων ταξινόμησης μιας κλάσης. Τα K διαγράμματα χρησιμοποιούν ένα παρακολουθητικό στατιστικό το οποίο απορρέει από την απόσταση ανάμεσα στην νέα παρατήρηση και στο όριο απόφασης που έχει υπολογιστεί από τον αλγόριθμο SVDD. Τα όρια ελέγχου των K διαγραμμάτων εγκαθιδρύονται και προσαρμόζονται από μία παράμετρο του SVDD αλγορίθμου. Η μελέτη των Sun και Tsung αποκαλύπτει ότι τα K διαγράμματα αποδίδουν καλύτερα από τα T^2διαγράμματα όταν τα δεδομένα αποκλίνουν από την κανονικότητα. Ο Kumar (2006) χρησιμοποίησε μια άλλη SVM τεχνική μιας τάξης για να κατασκευάσει εύρωστα K διαγράμματα μέσω κανονικοποιημένων στατιστικών και έδειξε ότι εκτός από τα εύκαμπτα μη-κανονικά δεδομένα, τα εύρωστα K διαγράμματα μπορούν να διαχειριστούν καλύτερα τα αυτοσυσχετισμένα δεδομένα της διαδικασίας. Πέρα από όλα αυτά, τα διαγράμματα ελέγχου που βασίζονται στην SVM τεχνική μίας τάξης, έχουν εφαρμοστεί κατάλληλα έτσι ώστε να ανιχνεύουν ανωμαλίες στις εφαρμογές δικτύωσης των υπολογιστών (Zhang *et al.*, 2007).

Αξίζει να σημειωθεί σε αυτό το σημείο ότι οι προαναφερθείσες μελέτες χρησιμοποιούν στατιστικά παρακολούθησης προερχόμενα από τη SVM τεχνική μίας τάξης, γεγονός που έχει σαν αποτέλεσμα η κατασκευή των διαγραμμάτων να μην απαιτεί την παραδοχή κάποιας κατανομής. Όμως ακόμα δεν έχει προταθεί ένας αποτελεσματικός τρόπος κατασκευής των ορίων ελέγχου που αποτελεί ένα από τα βασικά συστατικά των διαγραμμάτων ελέγχου.

6.2 Διαγράμματα ελέγχου βασισμένα στον SVDD αλγόριθμο:

Στην εργασία των Sukchotrat, Kim και Tsung (2010), περιγράφονται τα πιο σημαντικά μη παραμετρικά διαγράμματα ελέγχου.

6.2.1 Ο Αλγόριθμος SVDD

Είναι αλγόριθμος Μηχανής Υποστήριξης Διανυσμάτων (Support Vector Machine-SVM) είναι ένας από τους αλγόριθμους μάθησης με επίβλεψη οι οποίοι χρησιμοποιούνται σε προβλήματα παλινδρόμησης και ταξινόμησης. Οι SVMs χρησιμοποιούν γεωμετρικές ιδιότητες και αποκτούν ένα διαχωριστικό υπερεπίπεδο λύνοντας ένα πρόβλημα κυρτής βελτιστοποίησης που ταυτόχρονα ελαχιστοποιεί το σφάλμα γενίκευσης και μεγιστοποιεί το γεωμετρικό περιθώριο μεταξύ των τάξεων (Vapnik, 1998). Μη γραμμικά SVM μοντέλα μπορούν να κατασκευαστούν από συναρτήσεις πυρήνα όπως είναι για παράδειγμα οι γραμμικές, οι πολυωνυμικές και οι ακτινικές (radial basis) συναρτήσεις.

Ο SVDD αλγόριθμος αποτελεί μία μίξη της SVM τεχνικής και της περιγραφικής μεθόδου δεδομένων για την επίλυση προβλημάτων ταξινόμησης μίας τάξης. (Tax and Duin, 2004) Ο SVDD αλγόριθμος παρέχει ένα σύνορο-όριο υπερσφαίρας γύρω από τα δεδομένα.

Έστω ότι το a αποτελεί το κέντρο της υπερσφαίρας και $το\ \ R^2$ αποτελεί την ακτίνα της υπερσφαίρας. Έστω τώρα $x_i = [x_{i1}, x_{i2}, \dots, x_{ip}]^T$, i = 1,2,…N μια ακολουθία p-μεταβλητών παρατηρήσεων εκπαίδευσης. Τα SVDD σύνορα κατασκευάζονται με στόχο να ελαχιστοποιούν τον όγκο της υπερσφαίρας ενώ ταυτόχρονα μεγιστοποιούν τις παρατηρήσεις εκπαίδευσης που έχουν συλληφθεί από την υπερσφαίρα.

Το πρόβλημα περιγράφεται μαθηματικά ως εξής :
Ελαχιστοποίησε το

$$R^2 + C\sum_{i=1}^{N}\xi_i \quad (2)$$

με τον περιορισμό

$$\|x_i - a\|^2 \leq R^2 + \xi_i \quad (3)$$

όπου $\xi_i > 0$ είναι η «χαλαρή» μεταβλητή (slack variable) που επιτρέπει στο x να βρίσκεται έξω από την υπερσφαίρα. Το C ελέγχει την εξισορρόπηση μεταξύ του όγκου της υπερσφαίρας και των σφαλμάτων της λανθασμένης ταξινόμησης. Οι Tax και Duin (Tax and Duin, 2004) όρισαν την παράμετρο

$$f = \frac{1}{NC} \quad (4)$$

η οποία καθορίζεται από τον χρήστη-πειραματιστή, και αναπαριστά το κλάσμα των δεδομένων εκπαίδευσης τα οποία βρίσκονται εκτός των συνόρων απόφασης με το N να αναπαριστά τον αριθμό των παρατηρήσεων στόχου (target observations).

Για παράδειγμα, το 80% των σημείων των δεδομένων εκπαίδευσης υποτίθεται ότι είναι εντός των SVDD ορίων τα οποία κατασκευάστηκαν για f = 0.20. Όταν το f αυξάνεται από 0.20 σε 0.30 ο όγκος της υπερσφαίρας γίνεται μικρότερος αλλά το σφάλμα λανθασμένης ταξινόμησης στην κλάση-στόχος γίνεται μεγαλύτερο. Το πρόβλημα της ελαχιστοποιήσης που περιγράφεται παραπάνω από τη σχέση (2) μπορεί να επιλυθεί κάνοντας χρήση της παρακάτω Λαγκρανζιανής

$$L(R, a, a_i, \gamma_i, \xi_i) = R^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}a_i\{R^2 + \xi_i - (\|x_i - a\|^2)\} - \sum_{i=1}^{N}\gamma_i\xi_i \quad (5)$$

όπου $a_i, \gamma_i \geq 0$ είναι οι λαγκρανζιανοί πολλαπλασιαστές.

Υπολογίζοντας τις μερικές παραγώγους του L ως προς τα R, a, ξ_i και θέτοντάς τις ίσες με 0 έχουμε:

$$\sum_{i=1}^{N}a_i = 1 \quad (6)$$

$$a = \sum_{i=1}^{N}a_i\, x_i \quad (7)$$

$$a_i = C - \gamma_i \quad (8)$$

Αντικαθιστώντας τους περιορισμούς αυτούς στην σχέση (5), το πρόβλημα βελτιστοποίησης γίνεται:

$$L = \sum_i a_i \left(x_i \cdot x_j\right) - \sum_{ij} a_i\, a_j \left(x_i \cdot x_j\right) \quad (9)$$

Η λύση στο πρόβλημα μεγιστοποίησης, δηλαδή το σύνολο των a_i, i = 1,2,…N προκύπτει μεγιστοποιώντας τη σχέση (9) υπό τις συνθήκες:

$$0 \leq a_i \leq C$$

και

$$\sum_{i=1}^{N} a_i = 1$$

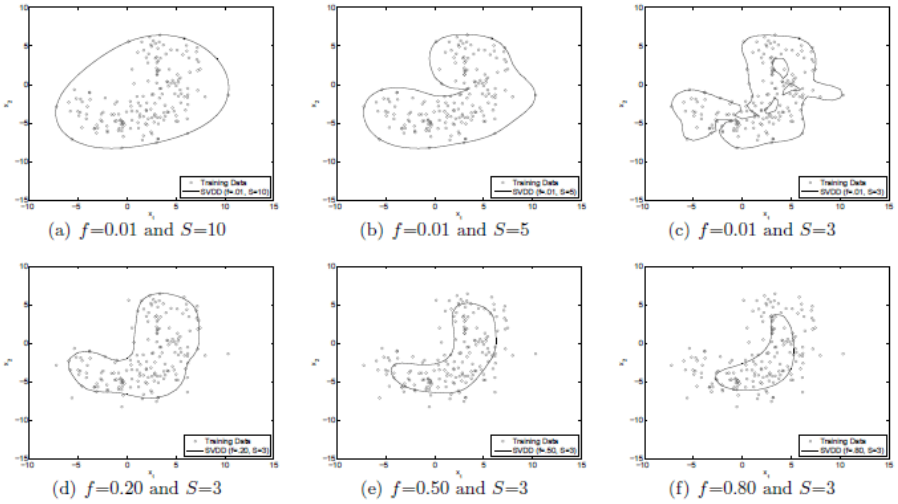
Ο αλγόριθμος SVDD μπορεί να παράγει πιο ευέλικτα όρια απόφασης αντικαθιστώντας το εσωτερικό γινόμενο με συναρτήσεις πυρήνα. Για παράδειγμα η παρακάτω Γκαουσιανή συνάρτηση πυρήνα μπορεί να αντικατασταθεί με το εσωτερικό γινόμενο της σχέσης (9)

$$K\left(x_i \cdot x_j\right) = \exp\left(-\frac{\left\|x_i - x_j\right\|}{S^2}\right) \quad (10)$$

όπου $S > 0$ είναι το εύρος του Γκαουσιανού πυρήνα που ελέγχει την πολυπλοκότητα του συνόρου του SVDD. Δεδομένου ενός σημείου από τα δεδομένα δοκιμής z, το D^2 που υπολογίζει την απόσταση μεταξύ του z και του κέντρου και το a μπορούν να υπολογιστούν από την ακόλουθη εξίσωση:

$$D^2 = K(z \cdot z) - 2\sum_i a_i\, K(z \cdot x_i) - \sum_{ij} a_i a_j\, K\left(x_i \cdot x_j\right) \quad (11)$$

Για την επίδειξη των ορίων ελέγχου του SVDD κατασκευάστηκε ένα σύνολο δεδομένων σε σχήμα μπανάνας χρησιμοποιώντας Matlab κώδικα διαθέσιμο από τα PRTools (Duin *et al.*, 2007). Τα όρια ελέγχου με διαφορετικές τιμές στις παραμέτρους (f *και* S) στον αλγόριθμο SVDD κατασκευάστηκαν από 180 εντός-ελέγχου παρατηρήσεις εκπαίδευσης (Φάση I δεδομένα). Το Σχήμα 1 δείχνει διαφορετικά SVDD όρια εμπεδωμένα σε δισδιάστατα διαγράμματα των δεδομένων της Φάσης I. Αν δώσουμε την ίδια τιμή στο f= 0.01 το σχήμα των ορίων ελέγχου γίνεται πιο ομαλό για μεγαλύτερο S. Επίσης από το Σχήμα 1 έχουμε για την ίδια τιμή του *S* (*S*=3) το σύνορο ελέγχου γίνεται πιο στενό στο κέντρο μάζας για μεγαλύτερο *f*.



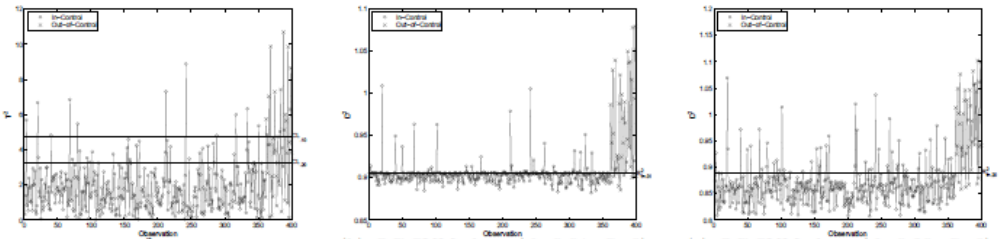
Σχήμα 1.: Σύνορα ελέγχου του SVDD που προκύπτουν από διαφορετικές τιμές των παραμέτρων f και S.

6.2.2 Υπάρχουσες μέθοδοι διαγραμμάτων ελέγχου βασισμένα στον SVDD αλγόριθμο.

Αρκετές μελέτες έχουν ενσωματώσει μεθόδους ταξινόμησης μίας κλάσης σε SPC προβλήματα. Οι Sun και Tsung, (2003) πρότειναν τα K διαγράμματα για να χειριστούν μη-κανονικά προβλήματα χρησιμοποιώντας τις αποστάσεις από τον πυρήνα οι οποίες υπολογίζονται με τον αλγόριθμο SVDD. Καθιέρωσαν και προσάρμοσαν τα όρια ελέγχου του K διαγράμματος χρησιμοποιώντας το f ή το C, μία από τις παραμέτρους του SVDD αλγορίθμου. Ο Kumar, (2006) πρότεινε τα εξομαλυσμένα K διαγράμματα που είναι παρόμοια με τα K διαγράμματα αλλά χρησιμοποιούν κανονικοποιημένες αποστάσεις πυρήνα. Τα διαγράμματα ελέγχου που βασίζονται σε μίας κλάσης SVM εφαρμόστηκαν για την ανίχνευση ανωμαλιών στα δίκτυα υπολογιστών. Τα διαγράμματα ελέγχου που βασίζονται σε SVM ταξινόμηση μίας κλάσης (OC-SVM) χρησιμοποιούνται για την ανίχνευση ανωμαλιών στα δίκτυα υπολογιστών (Zhang *et al.*, 2007). Τα διαγράμματα ελέγχου που βασίζονται σε SVM ταξινόμηση μίας κλάσης (OC-SVM) κατασκευάζονται με την σχεδίαση των στατιστικών παρακολούθησης (D^2) που υπολογίζουν 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 τα επίπεδα σφάλματος στα OC-SVM διαγράμματα ρυθμίζονται από το f. Μεγαλύτερες τιμές του f τείνουν να αποφέρουν μεγαλύτερο ποσοστό σφαλμάτων τύπου I επειδή ο αλγόριθμος χρησιμοποιεί λιγότερα δεδομένα εκπαίδευσης εντός του συνόρου. Στα OC-SVM διαγράμματα ο καθορισμός της τιμής του f από τον χρήστη επηρεάζει όχι μόνο τον καθορισμό των ορίων ελέγχου αλλά και τον υπολογισμό του στατιστικού παρακολούθησης. Λόγω των μεγάλων διαφορών

στα διαγράμματα ελέγχου για διαφορετικό f , το f είναι ακατάλληλο για να εγκαθιδρύσει τα όρια ελέγχου στα OC-SVM διαγράμματα. Μια παρατήρηση που έχει ανιχνευτεί ως εκτός-ελέγχου (ή εντός-ελέγχου) μπορεί να μην ανιχνεύεται πια ως εκτός-ελέγχου (αντίστοιχα εντός-ελέγχου) για διαφορετικές τιμές του f. Σε αντίθεση, τα διαγράμματα T^2 χρησιμοποιούν την μεταβλητή ελέγχου a που είναι ανεξάρτητη από το στατιστικό ελέγχου T^2. Το ελλειπτικό σύνορο του T^2 περιλαμβάνει περισσότερες εκτός-ελέγχου παρατηρήσεις και αποφέρει ένα υψηλότερο ποσοστό σφάλματος τύπου I για μεγαλύτερο a. Επίσης, οι ίδιες τιμές των στατιστικών ελέγχου σχεδιάζονται στο διάγραμμα T^2 ανεξάρτητα από το a.


(a) T^2 chart (b) OC-SVM chart (f=0.01, S=3) (c) OC-SVM chart (f=0.20, S=3)

Σχήμα 2. - T^2 και OC-SVM διαγράμματα με τα στατιστικά και τα όρια ελέγχου που αντιστοιχούν στα (c) και (d) σύνορα ελέγχου του Σχήματος 1.

6.2.3 Νέα στρατηγική σχεδιασμού των OC-SVM διαγραμμάτων που βασίζονται στην Bootstrap τεχνική:

Λόγω των περιορισμών που προέκυπταν από τη χρήση των OC-SVM διαγραμμάτων στην αρχική τους μορφή προτάθηκαν τα D^2 διαγράμματα για να καθορίσουν τα όρια ελέγχου στα OC-SVM διαγράμματα. Τα όρια ελέγχου των διαγραμμάτων D^2 προσαρμόστηκαν βασισμένα σε μια ποσοστιαία τιμή υπολογισμένη από την μέθοδο Bootstrap. Η Bootstrap είναι μία από τις πιο διαδεδομένες μεθόδους επαναδειγματοληψίας, που παρέχει στατιστικές εκτιμήσεις όταν η κατανομή του πληθυσμού είναι άγνωστη (Efron and Tibshirani, 1993).

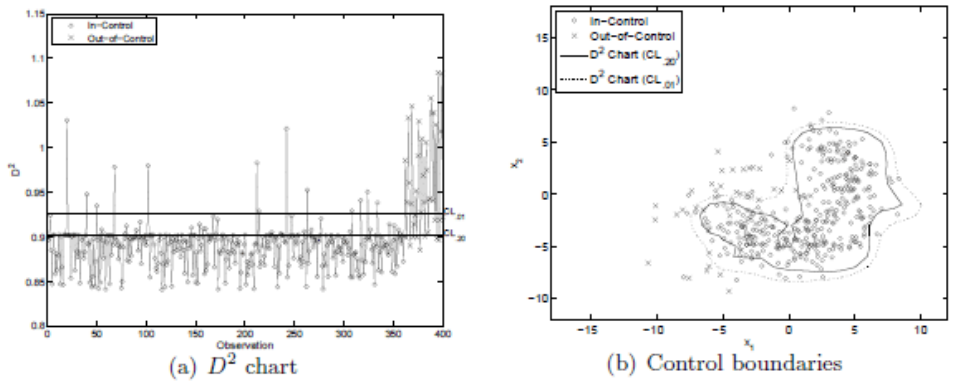
Στα παραδοσιακά διαγράμματα ελέγχου τα όρια ελέγχου καθορίζονται από την υποκείμενη κατανομή του στατιστικού ελέγχου με την καθορισμένη από τον χρήστη τιμή. Σε αντίθεση, η κατανομή του στατιστικού ελέγχου ενός D^2 διαγράμματος είναι άγνωστη εξαιτίας της μη-παραμετρικής φύσης του. Αυτό παρακίνησε τους Thuntee Sukchotrat, Seoung Bum Kim, Fugee Tsung να αναπτύξουν μια κατάλληλη μη-παραμετρική διαδικασία για να καθορίσουν το όριο ελέγχου. Πρώτον, οι τιμές του D^2 των παρατηρήσεων της Φάσης I μεγέθους N αποκτώνται μέσω του SVDD αλγορίθμου. Δεύτερον, παίρνουμε B bootstrap δειγματοληψίες και υπολογίζουμε τις ποσοστιαίες αξίες ενδιαφέροντος από κάθε bootstrap δείγμα μεγέθους N με

αντικατάσταση των D^2 τιμών των παρατηρήσεων της Φάσης I. Τελικά το όριο ελέγχου καθορίζεται με τον υπολογισμό του μέσου των B ποσοστιαίων τιμών.

Ακολουθεί αναλυτική περιγραφή της bootstrap μεθόδου η οποία καθορίζει τα όρια ελέγχου στο D^2 διάγραμμα.

i. Υπολογίστε τα D^2 στατιστικά των παρατηρήσεων της Φάσης I μεγέθους N χρησιμοποιώντας την εξίσωση (11) και πάρε B ανεξάρτητα bootstrap δείγματα. Έστω $D_{j1}^2, D_{j2}^2, \dots, D_{jN}^2$ ακολουθία από N σε πλήθος D^2 στατιστικά από το $j-οστό$ bootstrap δείγμα, για $j = 1, \dots, B$.
ii. Για κάθε bootstrap δείγμα, δοθέντων του καθορισμένου από τον χρήστη $\alpha, (0 < \alpha \leq 1)$ και των διατεταγμένων D^2 τιμών ($D_{j(1)}^2 < D_{j(2)}^2 < \dots < D_{j(N)}^2$), το $D_{j(i)}^2$ είναι η $i-οστή$ μεγαλύτερη τιμή από N σε πλήθος D^2 τιμές στο $j-οστό$ bootstrap δείγμα, όπου το i είναι ο στρογγυλοποιημένος αριθμός που προκύπτει από το γινόμενο $N \cdot \alpha$.
iii. Υπολογίστε το όριο ελέγχου (CL) υπολογίζοντας τον μέσο από τις $i-οστές$ μεγαλύτερες τιμές σε κάθε B bootstrap δείγμα: $CL = \sum_{j=1}^{B} D_{j(i)}^2 / B$.
iv. Παρακολουθώντας τις παρατηρήσεις της Φάσης II δήλωσε ποιες παρατηρήσεις είναι εκτός-ελέγχου αν οι αντίστοιχες D^2 τιμές ξεπερνούν το όριο ελέγχου.

Στο Σχήμα 3, φαίνεται το διάγραμμα D^2 και το αντίστοιχο σύνορο ελέγχου. Στο διάγραμμα D^2 οι 180 εντός-ελέγχου παρατηρήσεις χρησιμοποιήθηκαν για την εκτίμηση των ορίων ελέγχου καθώς και σχεδιάστηκαν 400 D^2 στατιστικά από τις παρατηρήσεις της Φάσης II. Στο Σχήμα 3 (b) απεικονίζεται το αντίστοιχο σύνορο ελέγχου που προκύπτει από το διάγραμμα D^2 στο Σχήμα 3 (a). Παρατηρείται ότι αν αυξήσουμε το α από 0.01 σε 0.20 ανιχνεύονται πιο πολλές παρατηρήσεις ως εκτός-ελέγχου.



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

Σχήμα 3. - Το διάγραμμα D^2 και το αντίστοιχο σύνορο ελέγχου.

6.3 Τα k-κοντινότερα γειτονικά δεδομένα περιγραφής και τα διαγράμματα ελέγχου που βασίζονται σε αυτά (kNNDD-Based Control Charts)

Ο αλγόριθμος SVDD περιλαμβάνει ένα πρόβλημα βελτιστοποίησης που απαιτεί υψηλό υπολογιστικό φορτίο κατά την διαδικασία εκπαίδευσης. Ο αλγόριθμος SVDD απαιτεί 4.06 ώρες σε μία υπολογιστική μηχανή για να εκπαιδεύσει το μοντέλο με 4000 διμεταβλητές παρατηρήσεις. Εξαιτίας του υψηλού υπολογιστικού κόστους τα D^2 διαγράμματα μπορεί να μην είναι αποτελεσματικά για μια διεργασία που χρειάζεται συχνή επανεκπαίδευση. Για να αντιμετωπιστεί το παραπάνω πρόβλημα, προτείνεται ένα νέο διάγραμμα ελέγχου που βασίζεται στην ταξινόμηση μίας κλάσης που καλείται K^2 διάγραμμα. Ο αλγόριθμος που χρησιμοποιείται σε ένα K^2 διάγραμμα απαιτεί περίπου 5.42 δευτερόλεπτα αντίστοιχα για ένα μοντέλο αποτελούμενο από 4000 διμεταβλητές παρατηρήσεις εκπαίδευσης. Τα K^2 διαγράμματα ελέγχου βασίζονται σε μία μέθοδο περιγραφής των k-κοντινότερων γειτονικών δεδομένων, η οποία επιλύει προβλήματα ταξινόμησης μίας κλάσης υπολογίζοντας την τοπική πυκνότητα των δεδομένων χρησιμοποιώντας έναν αλγόριθμο πλησιέστερων γειτόνων (Breunig *et al.*, 2000; Tax, 2001).

6.3.1 Ο *kNNDD* αλγόριθμος

Έστω $NN_i(z)$ να είναι η i-οστή πλησιέστερη γειτονική παρατήρηση εκπαίδευσης στο σημείο z που χρειάζεται να ταξινομηθεί. Έστω V ο όγκος της υπερσφαίρας που περιέχει τις i πλησιέστερες γειτονικές παρατηρήσεις εκπαίδευσης και N το μέγεθος του συνόλου εκπαίδευσης. Η τοπική πυκνότητα του z καθορίζεται από την σχέση:

$$d(z) = \frac{i/_N}{V\|z - NN_i(z)\|} \qquad (12)$$

Όμοια, η τοπική πυκνότητα του $NN_i(z)$ μπορεί να καθοριστεί από την σχέση:

$$d\big(NN_i(z)\big) = \frac{i/_N}{V\|NN_i(z) - NN_i(NN_i(z))\|} \qquad (13)$$

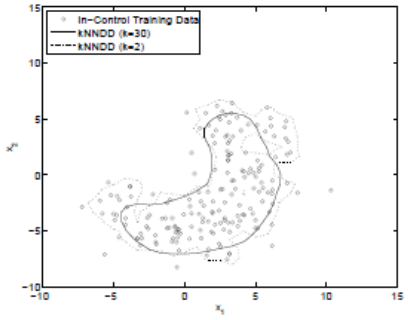
όπου $NN_i(NN_i(z))$ η $i-$οστόςπλησιέστερος γείτονας στο $NN_i(z)$ μέσα στο ίδιο σύνολο εκπαίδευσης. Ο αλγόριθμος $kNNDD$ ταξινομεί το z σαν την κλάση στόχο όταν η αναλογία της τοπικής πυκνότητας του z (12) προς την τοπική πυκνότητα του $NN_i(z)$ (13) είναι μεγαλύτερη ή ίση με το 1. Αυτό εξηγείται ως εξής:

$$\frac{d(z)}{d(NN_i(z))} = \frac{\|NN_i(z) - NN_i(NN_i(z))\|}{\|z - NN_i(z)\|} \geq 1 \qquad (14)$$

Για να γίνει ο αλγόριθμος πιο εύρωστος η σχέση (14) γίνεται:

$$\frac{\sum_{i=1}^{k}\big\|NN_i(z) - NN_i\big(NN_i(z)\big)\big\|}{\sum_{i=1}^{k}\|z - NN_i(z)\|} \geq 1 \qquad (15)$$

Το μέγεθος του πλησιέστερου γείτονα k, επηρεάζει την απόδοση του αλγορίθμου. Σύμφωνα με μελέτες το κατάλληλο εύρος τιμών για την μεταβλητή k στον αλγόριθμο $kNNDD$ είναι μεταξύ των τιμών 10 και 50 (Breunig *et al.*, 2000) .



Σχήμα 4. - Τα σύνορα ελέγχου του *kNNDD* με διαφορετικά *k* κατασκευασμένα από το σύνολο δεδομένων σχήματος μπανάνας.

6.3.2 K^2 διαγράμματα

Για την κατασκευή των K^2 διαγραμμάτων υπολογίζεται η απόσταση μεταξύ του z και των k πλησιέστερων παρατηρήσεων από την σχέση:

$$K^2 = \frac{\sum_{i=1}^{k} \|z - NN_i(z)\|}{k} \quad (16)$$

Οι τιμές K^2 χρησιμοποιούνται μετά σαν στατιστικά ελέγχου. Τα όρια ελέγχου ενός K^2 διαγράμματος καθορίζονται από την ποσοστιαία μέθοδο bootstrap όπως υποδεικνύεται στα διαγράμματα D^2.

Η αντίστοιχη ποσοστιαία διαδικασία bootstrap για τα διαγράμματα K^2 περιγράφεται αναλυτικά παρακάτω:

i. Υπολόγισε τα D^2 στατιστικά των παρατηρήσεων της Φάσης Ι μεγέθους N χρησιμοποιώντας την σχέση (11) και δημιούργησε B ανεξάρτητα δείγματα bootstrap. Έστω $K_{j1}^2, K_{j2}^2, \dots, K_{jN}^2$ μια ακολουθία από N στο πλήθος K^2 στατιστικών από το $j-οστό$ δείγμα bootstrap.

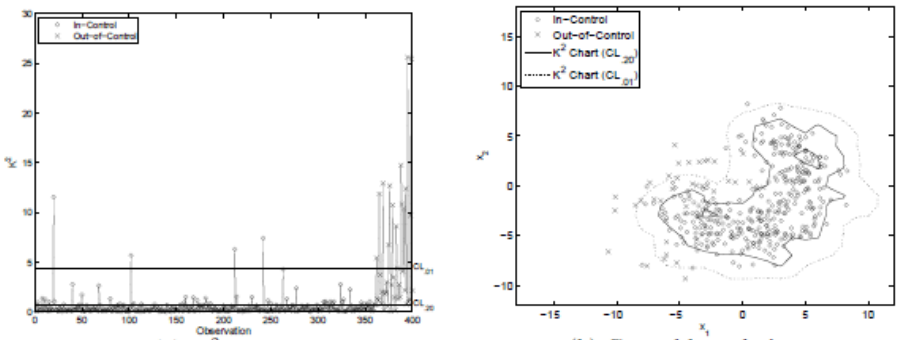
ii. Για κάθε bootstrap δείγμα όπου έχει δοθεί ένα καθορισμένο α από τον χρήστη $(0 < \alpha \leq 1)$ και οι διατεταγμένες K^2 τιμές, $(K_{j(1)}^2 < K_{j(2)}^2 < \cdots < K_{j(N)}^2)$, $\eta\, K_{j(i)}^2$ είναι η $i-οστή$ μεγαλύτερη παρατήρηση από τις N στο πλήθος K^2 τιμές στο $j-οστό$ δείγμα bootstrap, όπου i είναι το στρογγυλοποιημένο νούμερο από το γινόμενο $N \cdot \alpha$. Τα διαγράμματα ελέγχου γίνονται πιο ευαίσθητα όσο το α αυξάνεται.

iii. Υπολόγισε το όριο ελέγχου (CL) υπολογίζοντας τον μέσο από τις $i-οστές$ μεγαλύτερες παρατηρήσεις σε κάθε ένα από τα **B** δείγματα bootstrap

$$CL = \sum_{j=1}^{B} K_{j(i)}^2 \Big/ B$$

.

iv. Παρακολουθώντας τις παρατηρήσεις της Φάσης ΙΙ δήλωσε ποιες παρατηρήσεις είναι εκτός-ελέγχου αν οι αντίστοιχες K^2 τιμές ξεπερνούν το όριο ελέγχου.



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

Σχήμα 5. : Το διάγραμμα K^2 και το αντίστοιχο σύνορο ελέγχου

6.4 Μελέτη προσομοίωσης

6.4.1 Σχεδιασμός προσομοίωσης

Η μελέτη προσομοίωσης έχει σαν σκοπό την σύγκριση μεταξύ των K^2, D^2, T^2 και OC-SVM διαγραμμάτων. Τα δεδομένα δημιουργήθηκαν βάσει της διμεταβλητής κανονικής, της διμεταβλητής t και της διμεταβλητής γάμμα κατανομής και χρησιμοποιήθηκε ένα σύνολο δεδομένων σχήματος μπανάνας. Για τα D^2 και OC-SVM διαγράμματα χρησιμοποιήθηκε το πλάτος του γκαουσιανού πυρήνα, S=1 για την κανονική, την t και την γάμμα και S=3 για το σύνολο δεδομένων σχήματος μπανάνας. Για τα K^2 διαγράμματα η μεταβλητή k παίρνει την τιμή k=30.

Έστω μ_0 το δινυσμα μέσου και Σ_0 ο πίνακας συνδιακύμανσης των εντός-ελέγχου δεδομένων. Έστω $\mu_1 = \mu_0 + \delta$ το διάνυσμα μέσου των εκτός-ελέγχου δεδομένων. Το μέγεθος της μετατόπισης δ παρουσιάζεται από την μη κεντρική παράμετρο λ

$$\lambda = \sqrt{\delta^T \sum_0^{-1} \delta} \qquad (17)$$

Για να δημιουργήσουμε τα εκτός-ελέγχου δεδομένα για την διμεταβλητή κανονική, την διμεταβλητή t και την διμεταβλητή γάμμα κατανομή, θεωρήσαμε δύο τύπους μέσης μετατόπισης ($\lambda = 2$ και $\lambda = 3$). Για μια συγκεκριμένη τιμή του λ όλες οι μεταβλητές έχουν την ίδια μεταβολή και εδώ δεν λαμβάνουμε υπόψη μας την αλλαγή στην διακύμανση. Δημιουργήθηκαν δύο διαφορετικές γωνίες των σχημάτων

μπανάνας που συμβολίζουν τα εντός –ελέγχου δεδομένα και τα εκτός-ελέγχου δεδομένα. Τα σενάρια προσομοίωσης περιγράφονται συνοπτικά παρακάτω:

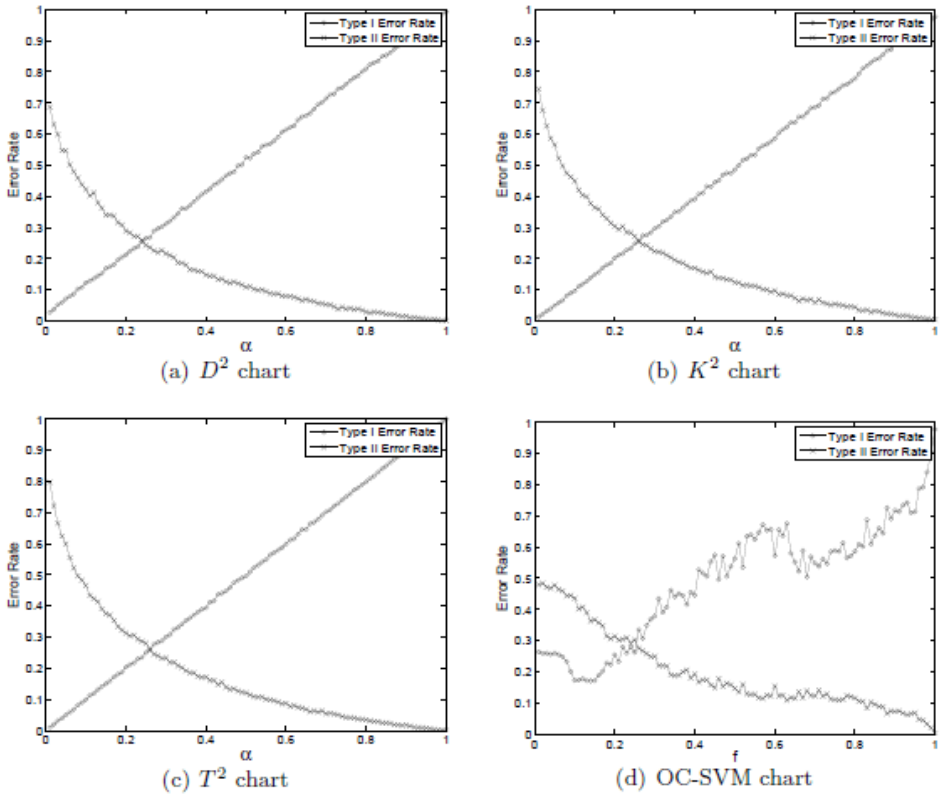
- $N_2,\ \lambda = 2$: Η περίπτωση της κανονικής κατανομής με μεσαία μεταβολή του μέσου. $\mu_0 = [0\ 0]$, $\Sigma_0 = \begin{bmatrix} 1 & 0.35 \\ 0.35 & 1 \end{bmatrix}$
- $N_2,\ \lambda = 3$: Η περίπτωση της διμεταβλητής κανονικής κατανομής με μεγάλη μεταβολή του μέσου. $\mu_0 = [0\ 0]$, $\Sigma_0 = \begin{bmatrix} 1 & 0.35 \\ 0.35 & 1 \end{bmatrix}$.
- $t_2(3),\ \lambda = 2$: Η περίπτωση της διμεταβλητής t κατανομής με 3 βαθμούς ελευθερίας, για μεσαία μεταβολή του μέσου.
- $t_2(3),\ \lambda = 3$: Η περίπτωση της διμεταβλητής t κατανομής με 3 βαθμούς ελευθερίας, για μεγάλη μεταβολή του μέσου.
- $Gam_2(1,1), \lambda = 2$: Η περίπτωση της διμεταβλητής γάμμα κατανομής με παραμέτρους σχήματος και κλίμακας ίσες με 1, για μεσαία μεταβολή του μέσου.
- $Gam_2(1,1), \lambda = 3$: Η περίπτωση της διμεταβλητής γάμμα κατανομής με παραμέτρους σχήματος και κλίμακας ίσες με 1,για μεγάλη μεταβολή του μέσου.
- Banana-shaped: Ένα σύνολο δεδομένων σε σχήμα μπανάνας, με δύο διαφορετικές γωνίες.

6.4.2 Όρια Ελέγχου

Σε αντίθεση με τα διαγράμματα OC-SVM τα οποία χρησιμοποιούν την παράμετρο f του αλγορίθμου SVDD για να προσαρμόσουν τα όρια ελέγχου, τα όρια ελέγχου των διαγραμμάτων D^2 και K^2 προσαρμόζονται από το εκατοστημόριο, που υπολογίζεται από την μέθοδο bootstrap. Τα διαγράμματα στα Σχήματα 6 και 7 δείχνουν πώς διαμορφώνονται τα επίπεδα σφάλματος τύπου I και II στα διαγράμματα K^2, D^2, T^2 και OC-SVM ανάλογα με του συντελεστές α και f. Από αυτά τα διαγράμματα συμπεραίνουμε ότι καθώς ο παράγοντας ελέγχου αυξάνεται, τα διαγράμματα ελέγχου παρουσιάζουν αυξημένα επίπεδα σφάλματος τύπου I και μικρότερα επίπεδα σφάλματος τύπου II. Η δυνατή αυτή θετική συσχέτιση μεταξύ του επιπέδου σφάλματος τύπου I και του παράγοντα ελέγχου είναι επιθυμητή. Τα D^2 και K^2 διαγράμματα ικανοποιούν την συνθήκη αυτή και σε κανονικές και σε μη-κανονικές περιπτώσεις. Τα διαγράμματα T^2 ικανοποιούν την συνθήκη αυτή μόνο σε περιπτώσεις κανονικής κατανομής. Σε κανονικές και σε μη-κανονικές περιπτώσεις τα διαγράμματα OC-SVM αποτυγχάνουν στο να παρουσιάζουν ισχυρή γραμμική συσχέτιση μεταξύ του επιπέδου σφάλματος τύπου I και του ελεγκτικού παράγοντα.

Τα επίπεδα σφάλματος τύπου I και II μπορεί να μην μπορούν να ελεγχθούν κατάλληλα από το f καθώς το μέγεθος των παρατηρήσεων στόχων αυξάνεται στα διαγράμματα OC-SVM. Παρατηρώντας το OC-SVM διάγραμμα για την περίπτωση $N_2,\ \lambda = 2$ χρησιμοποιώντας 300 και 400 παρατηρήσεις στόχου, διαπιστώνεται ότι τα

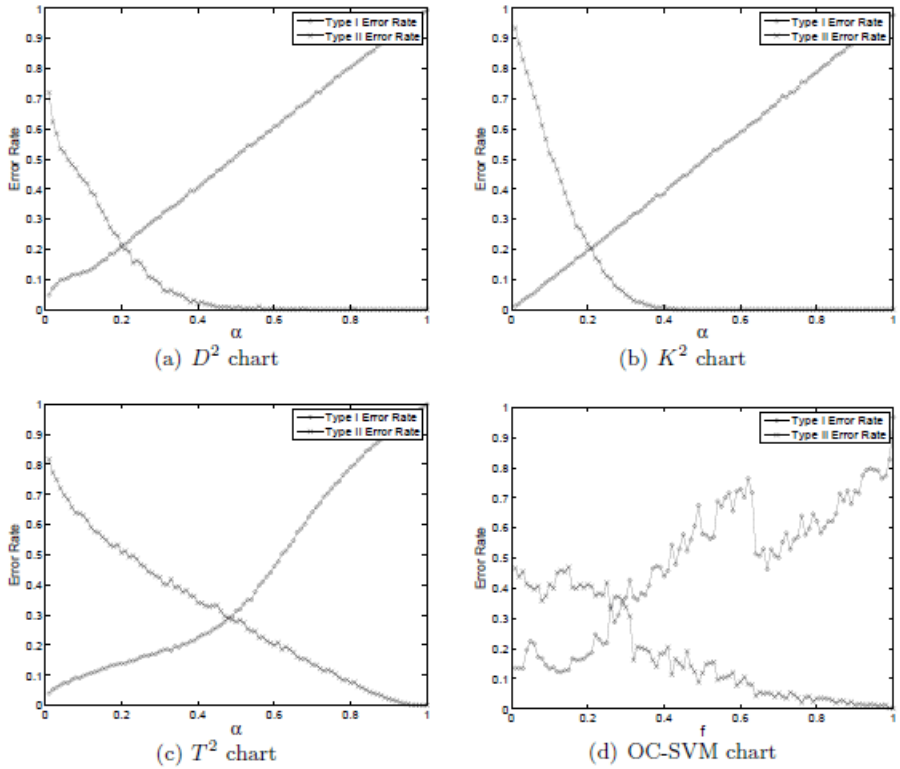
επίπεδα σφάλματος τύπου I και II είναι σταθερά για τις διάφορες τιμές του f. Το f όπως ορίστηκε στη σχέση (4) αντιπροσωπεύει το κλάσμα των δεδομένων στόχου εκτός του συνόρου απόφασης και είναι αντιστρόφως ανάλογο με τον συνολικό αριθμό των παρατηρήσεων στόχου. Συμπερασματικά, για μεγάλο αριθμό παρατηρήσεων στόχου, το f δεν παίζει σημαντικό ρόλο στην αλλαγή του συνόρου ελέγχου οδηγώντας σε σχετικά σταθερά επίπεδα σφαλμάτων τύπου I και II. Αυτό επιδεικνύει ότι το f είναι ακατάλληλη επιλογή ως παράγοντας ελέγχου στα OC-SVM διαγράμματα.



(a) D^2 chart (b) K^2 chart

(c) T^2 chart (d) OC-SVM chart

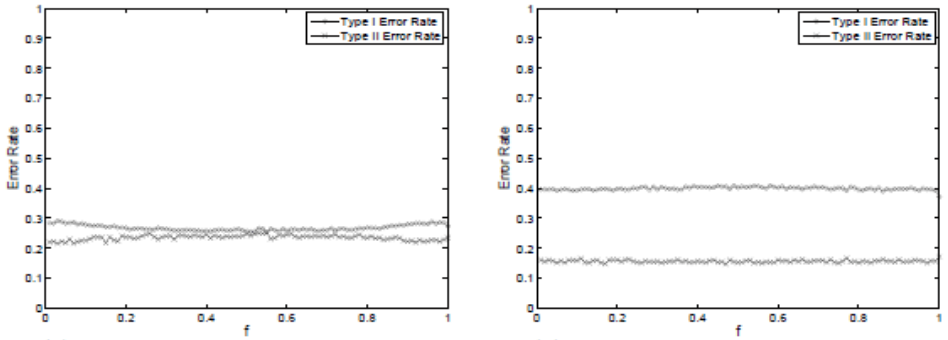
Σχήμα 6. - Μέσο ποσοστό σφάλματος τύπου I και τύπου II από τα διαγράμματα D^2, K^2, T^2, OC-SVM. (N_2 , $\lambda = 2$)



(a) D^2 chart (b) K^2 chart

(c) T^2 chart (d) OC-SVM chart

Σχήμα 7. - Μέσο ποσοστό σφάλματος τύπου I και τύπου II από τα διαγράμματα D^2, K^2, T^2, OC-SVM. $(Gam_2(1,1), \lambda = 2)$.



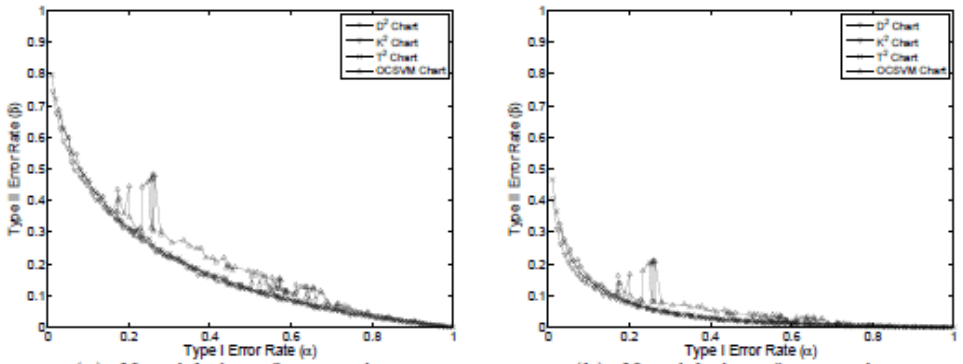
(a) Using 300 Phase I Observations (b) Using 400 Phase I Observations

Σχήμα 8. - Μέσα ποσοστά σφάλματος τύπου ι και τύπου II από τα διαγράμματα OC-SVM όταν ο αριθμός των παρατηρήσεων της φάσης I είναι μεγάλος. $(N_2\ , \lambda = 2)$

6.4.3 Συγκρίσεις Αποδόσεων

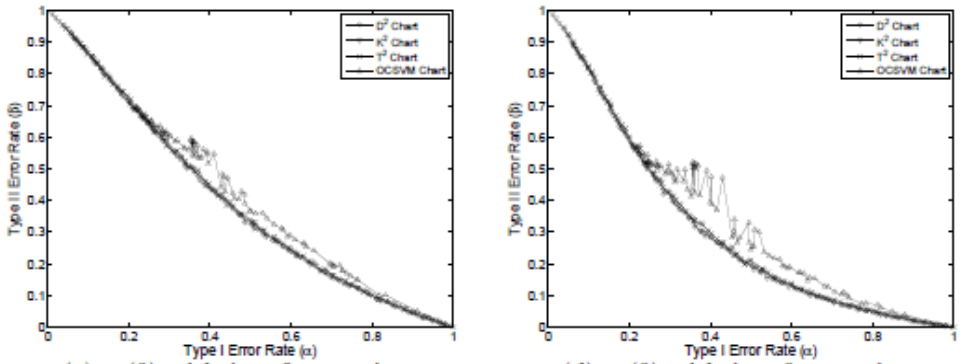
Το διάγραμμα ελέγχου που αποφέρει χαμηλότερο σφάλμα τύπου II θεωρείται καλύτερη μέθοδος αν το σφάλμα τύπου I είναι παρόμοιο. Τα διαγράμματα D^2 και K^2 παράγουν μικρότερου βαθμού σφάλματα τύπου II από το διάγραμμα T^2, δίνοντας παρόμοιου βαθμού σφάλματα τύπου I στα σενάρια των δεδομένων κατανομής γάμμα και σχήματος μπανάνας. Στις περιπτώσεις κανονικής και *t* κατανομής όλες οι μέθοδοι παρέχουν συγκρίσιμες αποδόσεις. Το εύρος των τυπικών σφαλμάτων 100 προσομοιώσεων είναι μεταξύ του 0.02 και του 0.08 για την κανονική κατανομή, την

t, και την γάμμα, ενώ από το διάγραμμα OC-SVM προκύπτουν μεγαλύτερα τυπικά σφάλματα μεταξύ 0.10 και 0.26. Τα OC-SVM διαγράμματα παράγουν ακανόνιστα σφάλματα τύπου II σε σχέση με τα σφάλματα τύπου I. Γι' αυτό είναι δύσκολο να συγκριθεί η απόδοσή τους με αυτή των άλλων διαγραμμάτων.



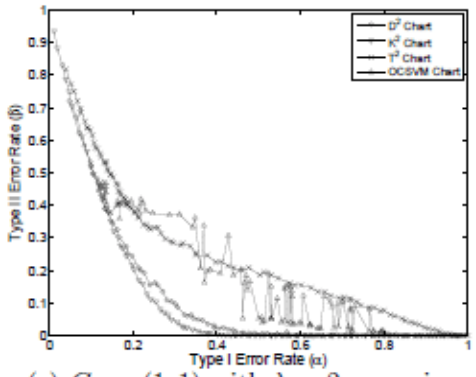
(a) N_2 with $\lambda = 2$ scenario

(b) N_2 with $\lambda = 3$ scenario

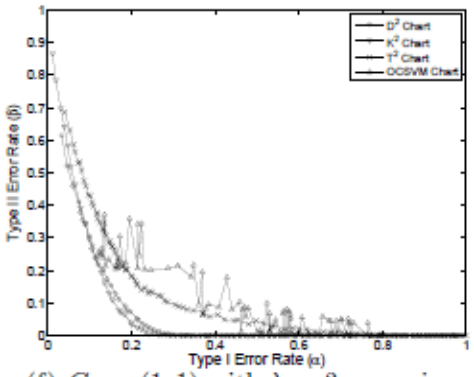


(c) $t_2(3)$ with $\lambda = 2$ scenario

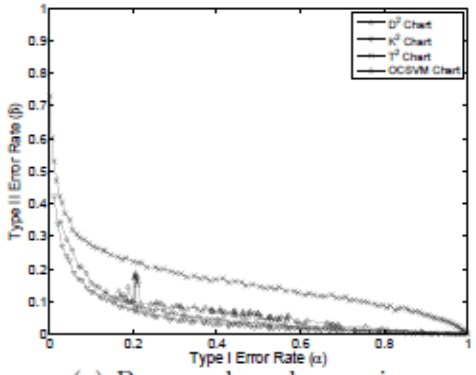
(d) $t_2(3)$ with $\lambda = 3$ scenario



(e) $Gam_2(1,1)$ with $\lambda = 2$ scenario



(f) $Gam_2(1,1)$ with $\lambda = 3$ scenario



(g) Banana-shaped scenario

Σχήμα 9. - Ποσοστά σφάλματος τύπου I και τύπου II των διαγραμμάτων D^2, K^2, T^2, OC-SVM υπό τα σενάρια προσομοιώσεων που αναπτύχθηκαν παραπάνω.

6.5 Εφαρμογή των διαγραμμάτων D2 και K2 στην Φάση I

Η ανάλυση στην Φάση I διαχωρίζει τα εντός-ελέγχου δεδομένα από το αρχικό σύνολο δεδομένων, το οποίο είναι ένα μείγμα από εντός και εκτός-ελέγχου δεδομένα, έτσι ώστε να καθορίσει αξιόπιστα όρια ελέγχου για τον έλεγχο μελλοντικών παρατηρήσεων. Η μελέτη προσομοίωσης έκανε σύγκριση της εφαρμοσιμότητας των διαγραμμάτων D^2 και K^2, για την υπάρχουσα μέθοδο στην Φάση I, η οποία αναδρομικά αφαιρεί τις παρατηρήσεις που ξεπερνούν τα όρια ελέγχου μέχρι να μην υπάρχουν καθόλου εκτός-ελέγχου παρατηρήσεις. Σε πείραμα με 200 αρχικές παρατηρήσεις, με παραμέτρους $\lambda = 2$ και $\lambda = 3$ τα διαγράμματα ελέγχου αφαίρεσαν τις παρατηρήσεις στις οποίες τα στατιστικά ξεπερνούσαν τα όρια ελέγχου και έτσι βρέθηκαν 20 παρατηρήσεις να είναι εκτός ελέγχου. Ανάλογα για τα D^2 και K^2 διαγράμματα, για την ανάλυση της Φάσης II, τα $100 \times (1-a)^{th}$ bootstrap εκατοστημόρια των D^2 και K^2 στατιστικών των αρχικών-ιστορικών δεδομένων χρησιμοποιήθηκαν σαν όρια ελέγχου στην ανάλυση της Φάσης I. Οι παρατηρήσεις που παρέμειναν ορίστηκαν ως εντός-ελέγχου. Οι παρατηρήσεις που ήταν εντός-ελέγχου αλλά αφαιρέθηκαν λανθασμένα ήταν τα σφάλματα τύπου I. Οι παρατηρήσεις που παρέμειναν ενώ ήταν εκτός-ελέγχου ήταν τα σφάλματα τύπου II.

Στην σύγκριση των διαγραμμάτων D^2 και K^2 με το T^2 στα επίπεδα σφαλμάτων τύπου I και II το συμπέρασμα είναι ότι οι αποδόσεις των D^2 και K^2 είναι συγκρίσιμες και είναι καλύτερες από το αναδρομικό T^2 κάτω από συνθήκες κανονικής και t κατανομής. Επειδή το διάγραμμα T^2 μπορεί να χειριστεί αποτελεσματικά τα δεδομένα πολυμεταβλητής κανονικής κατανομής, είναι κατάλληλο για τις περιπτώσεις ανάλυσης της Φάσης I στην κανονική κατανομή. Στις περιπτώσεις της κατανομής γάμμα και των δεδομένων σχήματος μπανάνας, τα διαγράμματα D^2 και K^2 παράγουν μικρότερα σφάλματα τύπου II από την αναδρομική T^2 μέθοδο. Αυτό σημαίνει ότι τα διαγράμματα D^2 και K^2 είναι αποτελεσματικές προσεγγίσεις για να χρησιμοποιηθούν στην ανάλυση της Φάσης I σε περιπτώσεις κανονικής και μη κανονικής κατανομής.

6.6 Συμπεράσματα

Στο κεφάλαιο αυτό παρουσιάστηκε ένα νέο είδος διαγραμμάτων, τα πολυμεταβλητά διαγράμματα ελέγχου τα οποία βασίζονται σε αλγορίθμους ταξινόμησης μίας κλάσης. Τα διαγράμματα D^2 και K^2 αποκτούν τα στατιστικά ελέγχου τους από τους αλγορίθμους SVDD και kNNDD. Τα όρια ελέγχου προέρχονται από τα bootstrap εκτιμώμενα ποσοστά των στατιστικών ελέγχου. Τα διαγράμματα ελέγχου τα οποία παρουσιάστηκαν παραπάνω, εξαιτίας της φύσης του να οδηγούνται από τα δεδομένα, μπορούν αποτελεσματικά να περιγράψουν την πραγματικότητα, τα χαρακτηριστικά των δεδομένων που παρακολουθούνται και απαιτούν ένα ελάχιστο σύνολο υποθέσεων για να κατασκευάσουν ένα διάγραμμα ελέγχου. Η συγκριτική μελέτη σύγκρισης των προσομοιωμένων δεδομένων έδειξε ότι οι αποδόσεις των διαγραμμάτων D^2 και K^2 είναι συγκρίσιμες με αυτές των T^2 διαγραμμάτων στην περίπτωση της κανονικής κατανομής. Όμως τα D^2 και K^2 υπερέχουν των T^2 διαγραμμάτων στις περιπτώσεις μη κανονικής κατανομής. Επιπρόσθετα, τα D^2 και K^2 υπερέχουν των T^2 διαγραμμάτων στις περιπτώσεις ανάλυσης προβλημάτων στη Φάση I. Στον πίνακα που ακολουθεί, παρουσιάζονται οι μέσες τιμές ποσοστών σφάλματος τύπου I (α) και τύπου II (β) του διαγράμματος D^2, του διαγράμματος K^2, και του T^2 στην Φάση I. Μέσα στις παρενθέσεις βρίσκονται οι μέσες τιμές των τυπικών σφαλμάτων

	D^2		K^2		T^2	
Scenarios	α	β	α	β	α	β
N_2, $\lambda = 2$	.1869	.3630	.1829	.3485	.1892	.3865
	(.0128)	(.0994)	(.0124)	(.1065)	(.0566)	(.1354)
N_2, $\lambda = 3$	.2124	.0800	.2183	.0825	.2137	.0915
	(.0103)	(.0674)	(.0097)	(.0561)	(.0664)	(.0810)
$t_2(3)$, $\lambda = 2$	.2219	.6830	.2144	.6790	.2248	.7165
	(.0119)	(.0932)	(.0122)	(.0970)	(.0426)	(.1071)
$t_2(3)$, $\lambda = 3$	.2038	.6290	.1995	.6375	.2069	.6230
	(.0132)	(.0949)	(.0126)	(.1013)	(.0444)	(.1436)
$Gam_2(1)$, $\lambda = 2$	.1996	.2410	.2101	.2825	.2089	.3250
	(.0171)	(.1307)	(.0140)	(.1196)	(.0596)	(.1969)
$Gam_2(1)$, $\lambda = 3$	.2204	.0380	.2208	.0715	.2335	.1200
	(.0100)	(.0556)	(.0116)	(.0905)	(.0781)	(.1482)
Banana-Shaped	.1751	.1680	.1771	.0865	.1791	.2380
	(.0118)	(.0886)	(.0127)	(.0721)	(.1026)	(.1211)

Σχήμα 10. - Μέσες τιμές ποσοστών σφάλματος τύπου I (α) και τύπου II (β) του διαγράμματος D^2, του διαγράμματος K^2, και του T^2 στην Φάση I.

7 ΚΕΦΑΛΑΙΟ – ΣΥΜΠΕΡΑΣΜΑΤΑ

Σε συνδυασμό με τις εφαρμοσμένες στατιστικές μεθόδους, το Data Mining είναι απαραίτητο εργαλείο για την εξαγωγή γνώσεων από τις βάσεις δεδομένων. Υπάρχουν δύο προσεγγίσεις για την εξόρυξη δεδομένων: Η Top-down ανάλυση, η οποία προσδιορίζει σχέσεις και τάσεις και η bottom-up ανάλυση, η οποία χρησιμοποιείται για την επιβεβαίωση ανακαλύψεων και την αξιολόγηση της ποιότητας των αποφάσεων. Στο Data-Mining δεν χρειάζονται υποθέσεις, υπάρχει η δυνατότητα εύρεσης προτύπων σε μεγάλες ποσότητες δεδομένων και χρησιμοποιούνται όλα τα διαθέσιμα δεδομένα. Η εξόρυξη δεδομένων παίζει σημαντικό ρόλο στην λήψη αποφάσεων μας εταιρίας. Τα πεδία στα οποία εφαρμόζεται είναι τα εξής: α. ανάλυση αποφάσεων μιας εταιρίας, β. ανάλυση αγοράς και διαχείριση, γ. εντοπισμός απάτης και διαχείριση ρίσκου.

Για την εξόρυξη δεδομένων αναλύθηκαν δύο μέθοδοι, η kNN και η SVM. Από τις επεκτάσεις της μεθόδου kNN, η προσέγγιση που χρησιμοποιεί ιδιότητες του συνόλου δεδομένων οδήγησε στην ανάπτυξη αποτελεσματικών μετασχηματισμών της μεθόδου kNN για τα προβλήματα ταξινόμησης. Οι παραλλαγές εξετάστηκαν ως προς την απόδοσή τους σε προβλήματα ταξινόμησης χρησιμοποιώντας δεδομένα από την UCI. Όλοι οι ταξινομητές που παρουσιάστηκαν έχουν καλή απόδοση, αν και κάποιοι ξεχωρίζουν σε μερικά σύνολα δεδομένων. Παρά το γεγονός ότι είναι πιο αργοί από τον kNN, είναι αξιόπιστοι υπό την έννοια του Δικτύου Αξιοπιστίας (NR). Περαιτέρω έλεγχος θα παρουσιάσει τα πλεονεκτήματα και τους περιορισμούς των ταξινομητών αυτών.

Η μέθοδος SVM βασίζεται στην στατιστική θεωρία εκμάθησης. Οι μηχανές υποστήριξης διανυσμάτων μπορούν να χρησιμοποιηθούν για την πρόβλεψη μελλοντικών δεδομένων μέσω διαδικασιών εκμάθησης. Υπάρχει η δυνατότητα να χρησιμοποιηθούν για την εκπαίδευση μιας ποικιλίας αναπαραστάσεων Τα τέσσερα βασικά χαρακτηριστικά των SVMs είναι η δυαδικότητα, οι πυρήνες, η κυρτότητα και η σπανιότητα. Μετά την διαδικασία εξόρυξης δεδομένων, ακολουθεί ο στατιστικός έλεγχος διεργασιών, οποίος πραγματοποιείται με την χρήση διαγραμμάτων ελέγχου. Διακρίνονται δύο βασικές κατηγορίες διαγραμμάτων ανάλογα με το είδος της μεταβλητής που περιγράφει ένα ποιοτικό χαρακτηριστικό του προϊόντος:

i. Διαγράμματα ελέγχου για συνεχή χαρακτηριστικά-μεταβλητές (control charts for variables)
ii. Διαγράμματα ελέγχου για διακριτά χαρακτηριστικά-ιδιότητες (control charts for attributes)

Αναφέρθηκαν επίσης οι διάφορες κατηγορίες διαγραμμάτων ελέγχου ανάλογα με την μορφή των δεδομένων και τον επιθυμητό έλεγχο. Έτσι, υπάρχουν τα διαγράμματα ελέγχου μεταβλητών, τα διαγράμματα ελέγχου ιδιοτήτων, τα αθροιστικά διαγράμματα ελέγχου, τα διαγράμματα ελέγχου με κινητούς μέσους και εκθετικά βάρη καθώς και τα μη παραμετρικά διαγράμματα ελέγχου. Τα τελευταία βασίζονται

σε αλγορίθμους ταξινόμησης μίας κλάσης. Τα διαγράμματα αυτά, μπορούν να περιγράψουν την πραγματικότητα, τα χαρακτηριστικά των δεδομένων και απαιτούν ένα ελάχιστο σύνολο υποθέσεων για να κατασκευάσουν ένα διάγραμμα ελέγχου.

Βιβλιογραφία

1. Abidin, T. And perizzo, W. SMART-TV: A Fast and scalable Nearest Neighbor Based Classifier for Data Mining. Proceedings of ACM SAC-06, Dijon, France, April 23-27, 2006. ACM Press, New York, NY, pp. 536-540
2. M. A. Aizerman, E. M. Braverman, and L. I. Rozono´er. Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control, 25:821–837, 1964.
3. Antal van den Bosch Feature Tranformation Through Rule induction: A Case Study with the k-NN Classifier. In J. Fürnkrantz(Ed.), Proceedings of the ECML/PKDD 2004 Workshop on Advances in Inductive Rule Learning, Pisa, Italy,September 2004, pp. 1-16
4. N. Aronszajn. Theory of reproducing kernels. Trans. Amer. Math. Soc., 686:337–404, 1950.
5. Aydin, T. And Guvenir, H. A. Modeling interestingness of Streaming Classification Rules as a Classification Problem. Lecture Notes in Computer Science, Springer Berlin/Heidelberg, ISSN 1611-349 (Online) Volume 3949-2006
6. Bakir, S., (2006) distribution-free quality control charts based on signed-rank-like statistics. *Communications in Statistics: Theory and Methods*, 35, 743-757.
7. Olivier Bousquet, Stephane Boucheron, and Gabor Lugosi, "Introduction to Statistical Learning Theory".
8. 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.
9. Burges B.~Scholkopf, editor, "Advances in Kernel Methods--Support Vector Learning". MIT press, 1998.
10. Burges C., "A tutorial on support vector machines for pattern recognition", In "Data Mining and Knowledge Discovery". Kluwer Academic Publishers, Boston, 1998, (Volume 2).
11. C. Cortes and V. Vapnik. Support vector networks. Machine Learning, 20:273 – 297,1995
12. Chakraborti, S., Van Der Laan, P. and Bakir, S.T. (2001) Nonparametric control chart: an overview and some results. *Journal of Quality Technology*, 33 (3), 304-315.
13. Chinnam, R.B. (2003) Support vextor machines for recognizing shifts in correlated and other manufacturing processes. *International Journal of Production Research*, 40 (17), 4449-4466.
14. Cook, D.F. and Chiu, C.C. (1998) Using radial basis function neural networks to recognize shifts in corre;ated manufacturing process parameters. *IIE Transactions*, 30 (3), 227-234.

15. Nello Cristianini and John Shawe-Taylor, “An Introduction to Support Vector Machines and Other Kernel-based Learning Methods”, Cambridge University Press, 2000.
16. Domeniconi, C., and Yan, B. On Error Correlation and accuracy of Nearest Neighbor Ensemble Classifiers Proceedings of the SIAM International Conference on Data Mining, Newport Beach, California, April 21-23, 2005
17. Duda R. and Hart P., "Pattern Classification and Scene Analysis", Wiley, New York 1973.
18. Efron, B. And Tibshirani, R. (1993) An Introduction to the Bootstrap. Chapman & Hall/ CRC, Boca Raton, FL
19. http://www.enm.bris.ac.uk/teaching/projects/2004_05/dm1654/kernel.htm
20. Theodoros Evgenuiu and Massimilliano Pontil, Statistical Learning Theory: a Primer 1998.
21. Hawkins D.M. (1991). A Fast, Accurate Approximation of Average Run Lengths of CUSUM Control Charts. Journal of Quality Technology, 24, 37-42
22. N. Heckman. The theory and application of penalized least squares methods or reproducing kernel hilbert spaces made easy, 1997.
23. Hendickx, I. And Antal van den Bosch. Maximum-Entropy Parameter Estimation for the k-nn Modified Value-Difference Kernel. Proceedings of the 16th Belgian-Dutch Conference on Artificial Intelligence, Groningen, The Netherlands, 2004
24. Hotelling, H. (1947) Multivariate quality control in *Techniques of Statistical Analysis*, Eisenhart, C., Hastay, M.W., and Wills, W.A. (eds), McGraw-Hill, New York, NY, pp. 111-184.
25. Hu, J., Gunger, G. And tuv, E. (2007) Tuned artificial contrasts to detect signals. *International Journal of Production Research*, 23 (1), 5527-5534.
26. Hwang, W.Y., Runger, G. and Tuv, E. (2005) Multivariate statistical process control with artificial contrasts. *IIE Transactions*, 39 (6), 659-669.
27. Khan, M., Ding, Q. And Perizzo, W. k-Nearest Neighbors Classification of Spatial Data Streams using P-trees. Proceedings of the PAKDD, 2002, pp. 517-528
28. Kim, S.H., Alexopoulos, C., Tsui, K.L. and Wilson, J.R. (2007) A distribution-free tabular CUSUM chart for autocorrelated data. *IIE Transactions*, 39 (3), 317-330.
29. Koukouvinos Ch. (2008). Statistical Process Control
30. Kumar, S., Choudhary, A.K., Kumar, M., Shankar, R. and Tiwari, M.K. (2006) Kernel distance-based robust support vector methods and its application in developing a robust K-chart. *International Journal of Production Research*, 44(1), 77-96.
31. J.P.Lewis, Tutorial on SVM, CGIT Lab, USC, 2004.
32. Liu, R.Y., Singh, K. and Teng, J.H. (2004) DDMA-charts: nonparametric multivariate moving average control charts based on data depth. *Allgemeines Statistisches Archiv*, 88(2), 235-258.

33. Lowry, C.A. and Montgomery, D.C. (1995) A review of multivariate control charts. *IIE Transactions*, 27(6), 800-810.
34. Lucas J. M. and M.S. Saccucci (1990). Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements. *Technometrics*, 32, 1-29.
35. Mainar-Ruiz, G. And Juan Carlos Pérez-Cortes Approximate Nearest Neighbor Search Using a Single Space-filling Curve and Multiple Representations of the Data Points. Proc. 18th International Conference on Pattern Recognition (ICPR 2006), 20-24 august 2006, Hong Kong, China:502-505
36. Mason, R.L. and Young, J.C. (2002) *Multivariate Statistical Process Control with Industrial Applications*. American Statistical Association and Society for Industrial and Applied Mathematics, Philadelphia, PA.
37. Tutorial slides by Andrew Moore. Http://www.cs.cmu.edu/~awm
38. Tom Mitchell, Machine Learning, McGraw-Hill Computer science series, 1997.
39. Monero-Seco, F., *Micó*, L. And Oncina, J., (2003). A modification of the LAESA Algorithm for approximated k-NN classification. *Pattern Recognition Letters*, **24** (1-3), 47-53.
40. Montgomery, D.C. (2005) *Introduction to Statistical Quality Control*, fifth edition. Wiley, New York, NY.
41. E. Osuna, R. Freund, and F. Girosi. An improved training algorithm for support vector machines. In J. Principe, L. Gile, N. Morgan, and E. Wilson, editors, Neural Networks for Signal Processing VII — Proceedings of the 1997 IEEE Workshop, pages 276 – 285, New York, 1997. IEEE.
42. Osuna E., Freund R., and Girosi F., "Support Vector Machines: Training and Applications", A.I. Memo No. 1602, Artificial Intelligence Laboratory, MIT, 1997.
43. Page, E. (1961). Cumulative Sum Control Charts. *Technometrics*, 3, 1-9.
44. Qiu, P. (2008) Distribution-free multivariate process control based on log-linear modeling. *IIE Transactions*, 40(7), 664-677.
45. Roberts S. W. (1959). Control Chart Tests Based on Geometric Moving Averages. *Technometrics*, 42, 97-102.
46. David M Skapura, Building Neural Networks, ACM press, 1996
47. Smith, A.E. (1994) *X* and *R* control chart interpretation using neural computing. *International Journal of Production Research*, 32(2), 309-320.
48. A. J. Smola. Regression estimation with support vector learning machines. Master's thesis, Technische Universität München, 1996
49. M. O. Stitson and J. A. E. Weston. Implementational issues of support vector machines. Technical Report CSD-TR-96-18, Computational Intelligence Group, Royal Holloway, University of London, 1996.
50. Stoumbos, Z.G., Reynolds, M.R., Ryan, T.P. and Woodall, W.H. (2000) The state of statistical process control as we proceed into the 21st century. *Journal of the American Statistical Association*, 95, 992-998.

51. T. Sukchotrat, S.-B. Kim, F. Tsung, (2010). One-class classification-based control charts for multivariate process monitoring. IIE Transactions, **42** (2), 107- 120**.**)
52. Sun, R. and Tsung, F. (2003) A kernel-distance-based multivariate control chart using support vector methods. *International Journal of Production Research*, 41(13), 2975-2989.
53. Tax, D.M.J. (2001) One-class classification: concept-learning in the absence of counter-examples. PhD thesis, Delf University of Technology, Netherlands.
54. Tax, D.M.J. and Duin, R.P.W. (2004) Support vector data description. *Machine Learning*, 54(1), 45-66.
55. Trafalis T., "Primal-dual optimization methods in neural networks and support vector machines training", ACAI99.
56. UCI Machine Learning Repository, available on line at the University of California, Irvine http://www.ics.uci.edu/~mlearn/MLSummary.html
57. V. Vapnik. The Nature of Statistical Learning Theory. Springer, N.Y., 1995. ISBN 0-387-94559-8.
58. V. Vapnik, S. Golowich, and A. Smola. Support vector method for function approximation, regression estimation, and signal processing. In M. Mozer, M. Jordan, and T. Petsche, editors, Advances in Neural Information Processing Systems 9, pages 281– 287, Cambridge, MA, 1997. MIT Press.
59. Vapnik V., ”Statistical Learning Theory”, Wiley, New York, 1998.
60. Vapnik, V., Estimation of Dependencies Based on Empirical Data. Empirical Inference Science: Afterword of 2006, Springer, 2006
61. Vapnik, V.N. (1998) *Statistical Learning Theory*. Wiley, New York, NY.
62. Veropoulos K., Cristianini N., and Campbell C., "The Application of Support Vector Machines to Medical Decision Support: A Case Study", ACAI99
63. Wang, H. And Bell, D. Extended k-Nearest Neighbours Based on Evidence Theory. The Computer Journal, Vol. 47 (6) Nov. 2004, pp. 662-672.
64. Wikipedia Online. http://en.wikipedia.org/wiki
65. Woodall, W.H. (2000) Controversies and contradictions in statistical process control. *Journal of Quality Technology*, 32(4), 341-350.
66. Yu, K. And Ji, L. Karyotyping af Comparative Genomic Hybridization Human Metaphases Using Kernel Nearest-NEighbor Algorithm, Cytometry, 48, 202-208, 2002
67. Zhang, H. and Albin, S. (2007) Determining the number of operational modes in baseline multivariate SPC data. *IIE Transactions*, 39(12), 1103-1110.
68. Zhang, Z., Zhu, X. and Jin, J. (2007) SVC-based multivariate control charts for automatic anomaly detection in computer networks. in *Proceedings of the Third International Conference on Autonomic and Autonomous Systems*.