ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Συγκριτική Μελέτη Μηχανισμών Εκτίμησης Ελλιπούς Πληροφορίας σε Ασύρματα Δίκτυα Αισθητήρων

Αιμιλία Β. Αργυροπούλου

Επιβλέποντες: **Ευστάθιος Χατζηευθυμιάδης,** Επίκουρος Καθηγητής
Βασίλειος Παπαταξιάρχης, Υποψήφιος Διδάκτωρ

ΑΘΗΝΑ

Δεκέμβριος 2011

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Συγκριτική Μελέτη Μηχανισμών Εκτίμησης Ελλιπούς Πληροφορίας σε Ασύρματα Δίκτυα Αισθητήρων

Αιμιλία Β. Αργυροπούλου
Α.Μ.: Μ1046

ΕΠΙΒΛΕΠΟΝΤΕΣ: **Ευστάθιος Χατζηευθυμιάδης,** Επίκουρος Καθηγητής
Βασίλειος Παπαταξιάρχης, Υποψήφιος Διδάκτωρ

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ: **Ευστάθιος Χατζηευθυμιάδης,** Επίκουρος Καθηγητής
Αθανασία Αλωνιστιώτη, Λέκτορας

Νοέμβριος 2011

ΠΕΡΙΛΗΨΗ

Ο κινητός υπολογισμός είναι ένας κλάδος που αναπτύσσεται ταχύτατα και έχει εφαρμογές σε πολλούς τομείς της ανθρώπινης δραστηριότητας. Τα ασύρματα δίκτυα αισθητήρων χρησιμοποιούνται κατά βάση στον κινητό υπολογισμό για την συλλογή των πληροφοριών. Η μείωση των διαστάσεων είναι μια τεχνική που εφαρμόζεται ευρέως κατά τη συλλογή των δεδομένων για να ελαττωθεί το αποθηκευτικό και υπολογιστικό τους κόστος, χωρίς μεγάλη απώλεια πληροφορίας. Η τεχνική που εξετάζεται στην παρούσα διπλωματική εργασία είναι η Ανάλυση Κύριων Συνιστωσών που μειώνει στο μισό τις αρχικές διαστάσεις των δεδομένων. Επίσης, τα ασύρματα δίκτυα αισθητήρων εμφανίζουν συχνά διάφορες μορφές σφαλμάτων που συνεπάγεται την ελλιπή πληροφόρηση προς τις εφαρμογές που καλύπτουν. Για την αποφυγή των προβλημάτων που μπορεί να προκύψουν από ελλιπείς τιμές, μελετούνται στατιστικές μέθοδοι όπως η προεκβολή / παρεμβολή καθώς και αλγόριθμοι κατηγοριοποίησης που προβλέπουν ελλιπείς τιμές σε ένα δείγμα δεδομένων. Οι αλγόριθμοι που εξετάζονται είναι οι: Decision Tree, C4.5, M5P, Decision Stump και RepTree. Τέλος, συγκρίνονται οι αποδόσεις των μεθοδολογιών και διεξάγονται συμπεράσματα για την αξιοπιστία τους βάσει των ποσοστών επιτυχίας που έχουν στην πρόβλεψη των μετρήσεων.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Αλγόριθμοι Πρόβλεψης
ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Κινητός Υπολογισμός, Ελαττωματικοί Αισθητήρες, Ανάλυση Κύριων Συνιστωσών, Προεκβολή, Κατηγοριοποίηση.

ABSTRACT

Mobile Computing constitutes a rapidly growing scientific area with many applications in everyday human activity. Mobile Computing exploits wireless sensor networks for the information collection. Data reduction techniques are applied in the collected data in order to reduce storage and computational cost, without much loss of information. The technique investigated in the current thesis is the Principal Component Analysis, which reduces the original size of the initial dataset. Further to the above, Wireless sensor networks often encounter various types of errors. As a result, incomplete data is supplied in the application systems. To face such phenomena, we performed statistical methods such as extrapolation / interpolation and classification algorithms which predict missing values in a sample data. The algorithms considered are: Decision Tree, C4.5, M5P, Decision Stump and RepTree. Finally, we present a comparison of the performance of the aforementioned methodologies. The conclusions are based on the success rates of the predicted values.

SUBJECT AREA: Prediction Algorithms

KEYWORDS: Mobile Computing, Faulty Sensors, Principal Component Analysis, Extrapolation, Classification.

Αφιερώνεται στην οικογένειά μου που με στηρίζει σε όλα μου τα βήματα.

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω ειλικρινώς τον κ. Ευστάθιο Χατζηευθυμιάδη καθώς και τον κ. Βασίλειο Παπαταξιάρχη για τις συμβουλές και τις υποδείξεις τους στην συγγραφή της διπλωματικής εργασίας.

Επίσης θα ήθελα να ευχαριστήσω την οικογένεια μου για την αμέριστη συμπαράστασή της.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΠΡΟΛΟΓΟΣ

Η παρούσα διπλωματική εργασία πραγματοποιήθηκε κατά τη διάρκεια του ακαδημαϊκού έτους 2010-2011, στα πλαίσια της φοίτησής μου στο Μεταπτυχιακό Πρόγραμμα Σπουδών του τμήματος Πληροφορικής & Τηλεπικοινωνιών, με κατεύθυνση «Προηγμένα Πληροφοριακά Συστήματα», του Εθνικού & Καποδιστριακού Πανεπιστημίου Αθηνών.

Στόχος ήταν να μελετηθούν τα σφάλματα που προκύπτουν σε ένα ασύρματο δίκτυο αισθητών κατά τη συλλογή δεδομένων και να γίνει μια συγκριτική μελέτη των μηχανισμών εκτίμησης ελλιπούς πληροφορίας.

Στα πρώτα κεφάλαια, γίνεται μια βιβλιογραφική μελέτη και ερμηνεία των μεθοδολογιών που ακολουθήθηκαν για την διεξαγωγή των πειραμάτων. Επίσης αναφέρονται τα εργαλεία που χρησιμοποιήθηκαν για την υλοποίηση των τεχνικών. Στα δύο τελευταία κεφάλαια αναλύονται τα αποτελέσματα και τα συμπεράσματα των πειραμάτων. Χρειάστηκε να επαναληφτούν αρκετές φορές τα πειράματα, και με διαφορετικά σενάρια ώστε να μπορέσουμε να αξιολογήσουμε την αξιοπιστία κάθε μεθοδολογίας.

1 ΕΙΣΑΓΩΓΗ

1.1 Διάχυτος Υπολογισμός

Με το πέρασμα των χρόνων, οι άνθρωποι χρησιμοποιούν ολοένα και περισσότερες υπολογιστικές συσκευές στην καθημερινότητά τους. Παλιότερα, η αναφορά μιας υπολογιστικής συσκευής παρέπεμπε σε έναν προσωπικό υπολογιστή. Αυτό δεν ισχύει πλέον. Στη σήμερον ημέρα, σχεδόν όλοι διαθέτουν εξελιγμένα κινητά τηλέφωνα με λειτουργίες αντίστοιχες των υπολογιστών αλλά και μικροσυσκευές καθημερινής χρήσης στις οποίες έχουν ενσωματωθεί υπολογιστικές συσκευές ώστε να διευρυνθούν οι λειτουργίες τους. Οι αισθητήρες παρκαρίσματος που βοηθούν τον οδηγό στην αντίληψη του χώρου, τα κλιματιστικά που σταματούν την λειτουργία τους όταν δεν βρίσκεται κανείς στο χώρο είναι μόνο μερικά παραδείγματα από την καθημερινή ζωή. Καθώς οι υπολογιστικές συσκευές εισβάλουν στη ζωή των ανθρώπων προκύπτει η ανάγκη να έχουν μεγαλύτερη υπολογιστική ισχύ για να μπορούν να καλύψουν περισσότερες λειτουργίες και γίνονται σταδιακά μικρότερες για να είναι πιο εύχρηστες.

Ο διάχυτος υπολογισμός (pervasive computing / ubiquitous computing) είναι αποτέλεσμα της ταχέως αναπτυσσόμενης τεχνολογίας των υπολογιστών και γενικότερα της τάσης που επικρατεί στην ενσωμάτωση υλικού (hardware) και λογισμικού (software) σε προϊόντα καθημερινής χρήσης. Ο διάχυτος υπολογισμός υποστηρίζει την ιδέα ότι σχεδόν οποιαδήποτε συσκευή μπορεί ιδανικά να συνδεθεί με ένα άπειρο δίκτυο άλλων συσκευών. Ο στόχος του διάχυτου υπολογισμού, που συνδυάζει τρέχουσες τεχνολογίες δικτύων και ασύρματων επικοινωνιών, το διαδίκτυο καθώς και την τεχνητή νοημοσύνη, είναι να δημιουργήσει ένα περιβάλλον όπου η συνδεσιμότητα είναι ενσωματωμένη με τέτοιο τρόπο ώστε να είναι διακριτική και πάντα διαθέσιμη.

1.1.1 Βασικές αρχές διάχυτου υπολογισμού

Από τότε που πρωτοεμφανίστηκαν οι υπολογιστές, και η χρήση της τεχνολογίας άρχισε να μπαίνει στην καθημερινότητα των ανθρώπων, μπορούμε να πούμε πως έχουν σημειωθεί τρεις σημαντικές τάσεις (περίοδοι) στην τεχνολογία [1]. Η οριοθέτηση αυτών των περιόδων γίνεται με βάση την επίδραση που είχαν στις ζωές των ανθρώπων. Είναι σημαντικό να γίνει ξεκάθαρο ότι δεν έχει να κάνει με την ίδια την τεχνολογία αλλά με την σχέση τεχνολογίας - ανθρώπου.

Η πρώτη περίοδος ονομάζεται «Περίοδος Μεγάλων Συστημάτων Υπολογιστών» (Mainframe Era) και αναφέρεται στην εποχή όπου οι άνθρωποι δεν ήταν καθόλου εξοικειωμένοι με τους υπολογιστές και η χρήση τους γινόταν μόνο από ειδικούς. Ένας υπολογιστής χρησιμοποιούταν από πολλούς χρήστες. Η δεύτερη μεγάλη τάση είναι αυτή του προσωπικού υπολογιστή (Personal Computer Era). Σε αυτή την περίοδο, το πλήθος των ανθρώπων που χρησιμοποιεί προσωπικούς υπολογιστές ξεπερνά αυτό που χρησιμοποιεί κοινόχρηστους υπολογιστές. Οι χρήστες έχουν τον δικό τους υπολογιστή, περιέχει τα δικά τους αρχεία και αλληλεπιδρούν άμεσα με αυτόν. Όταν οι χρήστες χρησιμοποιούν τους υπολογιστές είναι απασχολημένοι και δεν κάνουν κάτι άλλο. Τώρα διανύουμε την τρίτη περίοδο, αυτή της διάχυτης πληροφόρησης (Ubiquitous Computing Era), όπου ένας άνθρωπος χρησιμοποιεί πολλούς υπολογιστές χωρίς καν να το αντιλαμβάνεται. Το κύριο χαρακτηριστικό είναι η ενσωμάτωση εκατομμυρίων υπολογιστών στο περιβάλλον, επιτρέποντας στην τεχνολογία να χάνεται στο παρασκήνιο.

Σύμφωνα με τον Mark Weiser [2], οι σημαντικότερες και πλέον χρήσιμες τεχνολογίες είναι αυτές που «εξαφανίζονται», αυτές δηλαδή που τις χρησιμοποιούμε ευρέως χωρίς να συνειδητοποιούμε καν την ύπαρξή τους. Συνεπώς η ουσία του διάχυτου υπολογισμού είναι δεν είναι η ελαχιστοποίηση του όγκου του υπολογιστή ή η φορητότητά του. Αυτά είναι ορισμένα από τα μεταβατικά βήματα προς την επίτευξη των πραγματικών δυνατοτήτων των υπολογιστικών συστημάτων. Η χρήση των υπολογιστών οφείλει να είναι ένα «αόρατο» και αναπόσπαστο κομμάτι της καθημερινότητας των ανθρώπων.

Ο κινητός υπολογισμός (mobile computing) στηρίζεται σε μεγάλο βαθμό στο κλάδο των κατανεμημένων συστημάτων (distributed systems) και αποτελεί υποσύνολο του διάχυτου υπολογισμού. Ως εκ τούτου, αρκετά από τα ερευνητικά πεδία που μελετούνται σχετικά με τον διάχυτο υπολογισμό είναι κοινά με αυτά του κινητού υπολογισμού. Άρα, η έρευνα που είναι σχετική με τον διάχυτο υπολογισμό συμπεριλαμβάνει εκτός από τα ερευνητικά πεδία του κινητού υπολογισμού και επιπλέον σημαντικά θέματα που περιγράφονται παρακάτω.

- Μοντελοποίηση και Χρήση Ευφυών Χώρων (Smart Spaces): Η ενσωμάτωση υπολογιστικής υποδομής στην κτηριακή υποδομή είναι αυτό που χαρακτηρίζεται σαν «ευφυής χώρος» και επιτρέπει την παρακολούθηση και τον έλεγχο των δύο διαφορετικών υποδομών. Ένας χώρος μπορεί να είναι μέρος ενός κτιρίου (μια αίθουσα συνεδριάσεων, ένας διάδρομος, κτλ.) ή μπορεί να είναι μια καθορισμένη ανοικτή περιοχή όπως ένα προαύλιο ή ένας υπαίθριος χώρος.
- Αορατότητα (Invisibility): Το ιδανικό που εκφράστηκε από τον M. Weiser είναι η πλήρης απόκρυψη των τεχνολογιών του διάχυτου υπολογισμού από την επίγνωση του χρήστη. Στην πράξη, μια λογική προσέγγιση σε αυτό το ιδανικό είναι η ελάχιστη απόσπαση της προσοχής των χρηστών ώστε να αλληλεπιδρούν σχεδόν σε υποσυνείδητο επίπεδο με το περιβάλλον διάχυτου υπολογισμού.
- Δυνατότητα Κλιμάκωσης (Localized Scalability): Καθώς αναπτύσσονται οι ευφυείς χώροι, η αλληλεπίδραση ανάμεσα στον εξοπλισμό του χρήστη και το περιβάλλον αυξάνεται σημαντικά. Αυτό έχει σοβαρές επιπτώσεις στο εύρος ζώνης και την ενέργεια των τερματικών συσκευών. Όπως είναι φυσικό η αύξηση των χρηστών περιπλέκει το πρόβλημα. Έτσι, η κλιμάκωση, υπό την ευρύτερη έννοια, είναι μία πτυχή που πρέπει να ληφθεί σοβαρά υπ’ όψιν στον διάχυτο υπολογισμό.
- Απόκρυψη Διαφορετικών Συνθηκών Περιβάλλοντος (Masking Uneven Conditioning): Η ομοιόμορφη διείσδυση του διάχυτου υπολογισμού στις υποδομές απέχει πολλές δεκαετίες μακριά καθώς εξαρτάται από πολλούς και διαφορετικούς παράγοντες. Στο μεσοδιάστημα, η ανάπτυξη της «ευφυΐας» των διαφορετικών περιβαλλόντων αναμένεται να έχει τεράστιες διαφορές. Αυτό το τεράστιο εύρος διαφορετικών επιπέδων «ευφυΐας» στο περιβάλλον αναιρεί την έννοια της αορατότητας (invisibility) στην διάχυτη πληροφόρηση, καθώς γίνεται αντιληπτή από τους χρήστες. Ένας τρόπος για να μειωθεί αυτή η διαφορετικότητα, είναι να αντισταθμίσει ο χώρος των προσωπικών υπολογιστών τα διαφορετικά επίπεδα ευφυΐας του περιβάλλοντος. Για παράδειγμα, ένα σύστημα που έχει την δυνατότητα (λειτουργικότητα) της αυτόματης αποσύνδεσης, μπορεί να αποκρύψει την αδυναμία ασύρματης κάλυψης ενός άλλου περιβάλλοντος. Με αυτό τον τρόπο, δεν επιτυγχάνουμε πλήρη αορατότητα αλλά μειώνεται η μεταβλητότητα των διαφορετικών περιβαλλόντων.

1.1.2 Χρησιμότητα

Η έννοια του διάχυτου υπολογισμού βρίσκει εφαρμογές σε πάρα πολλούς τομείς της καθημερινότητας. Ορισμένοι τομείς όπου εφαρμόζεται ο διάχυτος υπολογισμός είναι η υγειονομική περίθαλψη και η φροντίδα υγείας στο σπίτι, τα συστήματα μεταφορών, η παρακολούθηση των μεταβολών του περιβάλλοντος κ.ο.κ.

Η χρήση του διάχυτου υπολογισμού στην υγειονομική περίθαλψη προσφέρει πολλά πλεονεκτήματα τόσο στην παρακολούθηση της θεραπείας όσο και στη διαχείριση των ασθενών. Υπάρχουν ειδικοί ασύρματοι αισθητήρες, που χρησιμοποιούνται για ιατρικούς σκοπούς και μπορούν να τοποθετηθούν στο χέρι (ή σε κάποιο άλλο σημείο σώματος) του ασθενή, έτσι ώστε να γίνονται οι απαραίτητες ιατρικές μετρήσεις. Τα αποτελέσματα που συλλέγονται, αφορούν φυσιολογικά δεδομένα (π.χ. παλμούς καρδιάς, αρτηριακή πίεση, δείκτη μάζας σώματος) αλλά και δεδομένα σχετικά με το περιβάλλον που βρίσκεται ο ασθενής (π.χ. θερμοκρασία δωματίου). Όσον αφορά τη διοίκηση του νοσοκομείου, μπορεί εύκολα να διαχειρίζεται τα δεδομένα των ασθενών, «μαρκάροντας» τον κάθε ασθενή με ειδική επικάρπια ετικέτα όπου θα συλλέγονται πληροφορίες σχετικά με το μητρώο του, τα φάρμακα που λαμβάνει και πιθανόν κάποιες ιατρικές σημειώσεις. Με αυτό τον τρόπο μειώνεται ο κίνδυνος εσφαλμένης ταυτοποίησης ασθενούς ή ακόμα και λάθη στη θεραπεία του. Ένα σύστημα που έχει αναπτυχθεί για την ηλεκτρονική παρακολούθηση των ασθενών που υποφέρουν από χρόνιες παθήσεις είναι το CHRONIC [4]. Το σύστημα CHRONIC περιλαμβάνει τρία διαφορετικά κομμάτια: έναν οικιακό κόμβο για τον ασθενή, αισθητήρες για απομακρυσμένη παρακολούθηση και ένα Κέντρο Διαχείρισης Χρόνιας Περίθαλψης (CCMC). Από τον οικιακό κόμβο ο ασθενής μπορεί να επικοινωνήσει με διάφορους φορείς υγειονομικής περίθαλψης (π.χ. τηλεφωνικώς, διάσκεψης μέσω βίντεο). Το σύστημα παρέχει αισθητήρες μέτρησης για το αναπνευστικό σύστημα, το ηλεκτροκαρδιογράφημα και το σφυγμό. Το Κέντρο Διαχείρισης Χρόνιας Περίθαλψης αποτελείται από ένα τηλεφωνικό κέντρο και ένα σύστημα διαχείρισης ασθενών το οποίο ενσωματώνεται στη δικτυακή εφαρμογή. Η μονάδα του τηλεφωνικού κέντρου λειτουργεί με μη-ιατρικό προσωπικό το οποίο κρίνει τις ανάγκες του ασθενή. Οι συνέπειες της επιτυχούς χρήσης του συστήματος είναι η μείωση των δαπανών της υγειονομικής περίθαλψης με τη μείωση του αριθμού νοσηλευόμενων ασθενών και η αίσθηση ασφάλειας που θα νιώθει ο ασθενής στο σπίτι του καθώς οι πιθανές έκτακτες ανάγκες μπορούν να ανιχνευθούν σε ένα αρχικό στάδιο.

Πολλές εφαρμογές διάχυτου υπολογισμού ασχολούνται με την ανάπτυξη ευφυών συστημάτων μεταφοράς. Με την εγκατάσταση αισθητήρων στους δρόμους και την ανάπτυξη επικοινωνίας μεταξύ οδικού δικτύου και οχήματος μπορούν να αποφευχθούν αρκετά από τα δυστυχήματα που συμβαίνουν καθημερινά. Επιπροσθέτως, δίνεται η δυνατότητα αποφυγής της κυκλοφοριακής συμφόρησης και η ανεύρεση συντομότερης διαδρομής. Αρκετά υπολογιστικά συστήματα που αφορούν τα οχήματα έχουν ήδη καθιερωθεί όπως αισθητήρες στάθμευσης και συστήματα εντοπισμού θέσης (Global Position System - GPS). Μια εφαρμογή που μπορεί να διαχειριστεί αποτελεσματικά την κυκλοφοριακή συμφόρηση είναι η Intelligent Traffic Information Service (ITIS) [5]. Το ITIS συλλέγει GPS δεδομένα σε πραγματικό χρόνο από οχήματα στα οποία έχουν τοποθετηθεί αισθητήρες και με μια προεπεξεργασία των δεδομένων κατευθύνει αντίστοιχα τα οχήματα. Στη συνέχεια, χρησιμοποιούνται αυτά τα δεδομένα για την εκτίμηση της κυκλοφοριακής κατάστασης. Το ITIS είναι ένα κατανεμημένο σύστημα που στοχεύει στην επεξεργασία μαζικών δεδομένων πραγματικού χρόνου ώστε να παρέχει υπηρεσίες πληροφόρησης με ελαχιστοποίηση της καθυστέρησης.

Η παρακολούθηση των μεταβολών του περιβάλλοντος είναι ανάμεσα στις πιο εμφανείς εφαρμογές του διάχυτου υπολογισμού. Επιτρέπει τη συνεχή και σε πραγματικό χρόνο συλλογή δεδομένων σχετικά με την υγρασία, τη θερμοκρασία, τις χημικές συνθέσεις κ.ο.κ., με τη χρήση απομακρυσμένων – ασύρματων συσκευών. Με την επεξεργασία αυτών των δεδομένων εξάγονται συμπεράσματα για τη γεωργία, τη σεισμολογία, τη μόλυνση του περιβάλλοντος κ.λ.π. Ορισμένα συστήματα που έχουν αναπτυχθεί για την Παρατήρηση του Περιβάλλοντος και Πρόβλεψη (Environmental Observations and Forecasting Systems – EOFS) είναι το CORIE [33] το οποίο μελετά τις εκβολές του ποταμού Κολούμπια, το FLOODNET [31] που προειδοποιεί για τυχόν πλημμύρες στο Ηνωμένο Βασίλειο και το SECOAS [32] που παρακολουθεί την διάβρωση γύρω από μικρά νησιά που προορίζονται για αιολικά πάρκα.

Αυτοί είναι κάποιοι από τους τομείς που εφαρμόζεται ο διάχυτος υπολογισμός. Με την πάροδο του χρόνου βρίσκει εφαρμογές σε περισσότερους κλάδους και αναπτύσσονται νέες μεθοδολογίες που υποστηρίζουν τη διάχυση πληροφορίας. Ο διάχυτος υπολογισμός είναι το επόμενο υπολογιστικό περιβάλλον, όπου η πληροφόρηση και η επικοινωνία θα είναι διαθέσιμες προς όλους, ανά πάσα στιγμή και από οποιοδήποτε σημείο. Οι τεχνολογίες της πληροφόρησης και της επικοινωνίας θα είναι αναπόσπαστο μέρος της καθημερινότητας του ανθρώπου.

1.2 Συστήματα Αισθητήρων

1.2.1 Ασύρματα Δίκτυα Αισθητήρων (WSN)

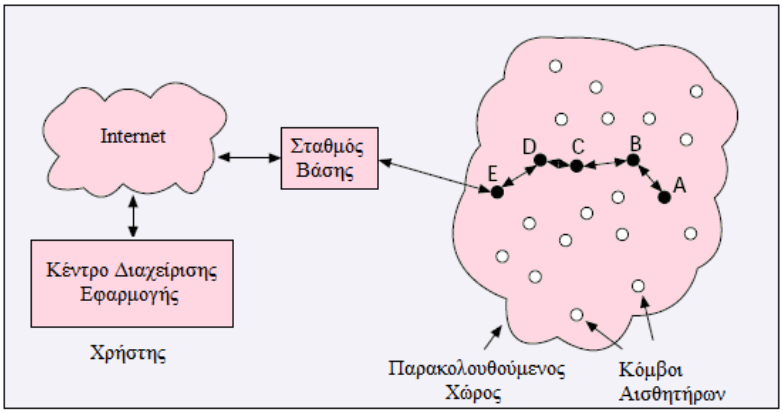
Τα δίκτυα αισθητήρων παρέχουν την κατάλληλη υποδομή ασύρματης επικοινωνίας ανάμεσα σε αισθητήρες που αναπτύσσονται για τη συλλογή και επεξεργασία δεδομένων συγκεκριμένων πεδίων εφαρμογών. Οι αισθητήρες είναι μικρές συσκευές που έχουν την ικανότητα της καταγραφής των μεταβολών των παραμέτρων, της επεξεργασίας των συλλεχθέντων δεδομένων και της επικοινωνία με το υπόλοιπο δίκτυο. Η δραστηριότητα των αισθητήρων μπορεί να είναι περιοδική ή σποραδική. Ένα παράδειγμα περιοδικής συλλογής δεδομένων είναι η καταμέτρηση περιβαλλοντικών μεταβολών όπως η θερμοκρασία και η υγρασία. Η ανίχνευση εισβολών σε σύνορα ή η ανίχνευση ότι η θερμοκρασία ενός φούρνου έχει ξεπεράσει το επιτρεπτό όριο είναι παραδείγματα σποραδικού τύπου [3].

Η υλοποίηση των «έξυπνων» αισθητήρων (smart sensors) πραγματοποιήθηκε με την ανάπτυξη της τεχνολογίας ασύρματων επικοινωνιών και με όλα τα πλεονεκτήματα που προσφέρει αυτή. Ένας «έξυπνος» αισθητήρας παρέχει πρόσθετη λειτουργικότητα από έναν απλό αισθητήρα. Μπορεί να ενσωματωθεί ευκολότερα σε ένα δικτυακό περιβάλλον επειδή μπορεί να επεξεργαστεί τα δεδομένα μέτρησης και να τα μετατρέψει στις αντίστοιχες μονάδες. Αυτό επιτυγχάνεται με μια συσκευή μνήμης που είναι ενσωματωμένη στον αισθητήρα και του δίνει τη δυνατότητα να αποθηκεύει πληροφορίες σχετικά με τη θέση (ταυτοποίηση μέτρησης), το εύρος των μετρήσεων, τη διόρθωση των δεδομένων [29].

Ένα δίκτυο αισθητήρων αποτελείται από ένα πλήθος αισθητήρων που είναι καταλλήλως τοποθετημένοι μέσα σε ένα παρακολουθούμενο χώρο. Η θέση των αισθητήρων δεν χρειάζεται να είναι προκαθορισμένη. Με αυτό τον τρόπο οι αισθητήρες έχουν την δυνατότητα να διασκορπιστούν τυχαία στον χώρο και να ερευνήσουν απρόσιτες περιοχές. Ωστόσο, αυτό απαιτεί σωστή διαχείριση χώρου από τους αλγόριθμους και τα πρωτόκολλα των δικτύων. Ένα άλλο χαρακτηριστικό των

«έξυπνων» αισθητήρων είναι ότι διαθέτουν ενσωματωμένους μικροεπεξεργαστές ώστε να εκτελούν απλούς υπολογισμούς στα δεδομένα που συλλέγουν. Αυτό έχει ως αποτέλεσμα την μετάδοση/ διαβίβαση μόνο των απαιτούμενων δεδομένων.

Η τυπική αρχιτεκτονική επικοινωνίας ενός ασύρματου δικτύου αισθητήρων είναι σχετικά απλή (Σχ.1.1). Οι κόμβοι των αισθητήρων διασκορπίζονται στον παρακολουθούμενο χώρο. Οι αισθητήρες είναι οργανωμένοι με τέτοιο τρόπο ώστε να συλλέγουν τα δεδομένα και να δρομολογούν τις μετρήσεις προς ένα σταθμό βάσης (base station/ sink). Τέλος, ο σταθμός βάσης επικοινωνεί, συνήθως, μέσω internet με το κέντρο διαχείρισης της εφαρμογής. Έτσι, οι χρήστες έχουν την δυνατότητα να παρακολουθήσουν και να επεξεργαστούν τις πληροφορίες που συλλέχθηκαν [6].



Εικόνα 1: Αρχιτεκτονική Επικοινωνίας Ασύρματου Δικτύου Αισθητήρων

1.2.2 Εφαρμογές Κινητού Υπολογισμού

Η κινητικότητα είναι κύριο χαρακτηριστικό της σύγχρονης ζωής. Οι άνθρωποι θέλουν να έχουν πρόσβαση σε πληροφορίες που αφορούν τόσο επαγγελματικά θέματα όσο και θέματα ψυχαγωγίας είτε βρίσκονται στη δουλειά τους είτε στο σπίτι τους είτε στο δρόμο. Με την ανάπτυξη της τεχνολογίας κινητών εφαρμογών αυτό είναι πλέον εφικτό. Για παράδειγμα, τα κινητά τηλέφωνα είναι αρκετά διαδεδομένα και δημοφιλή ώστε μπορούν να παρέχουν όλες τις απαραίτητες πληροφορίες στους ανθρώπους σε όποιο σημείο και αν βρίσκονται αυτοί. Για να υλοποιηθεί αυτό, οι κινητές εφαρμογές πρέπει να καλύπτουν απαιτήσεις πραγματικού χρόνου, κυρίως εάν πρόκειται να χρησιμοποιηθούν οι πληροφορίες για βραχυπρόθεσμες αποφάσεις.

Οι εφαρμογές κινητού υπολογισμού χρησιμοποιούνται για να επωφεληθούν πολυάριθμοι τομείς. Οι πιο βασικοί τομείς είναι η παρακολούθηση της κυκλοφορίας στο οδικό δίκτυο, η παρακολούθηση του περιβάλλοντος, η παρακολούθηση της υγείας ενός ασθενή κ.τ.λ. (όλοι αυτοί οι τομείς περιγράφηκαν αναλυτικότερα στην ενότητα 1.1.2.) Έτσι λοιπόν, με την χρήση εφαρμογών κινητού υπολογισμού θα μπορούν εύκολα οι άνθρωποι να παίρνουν απαντήσεις σε απλά και καθημερινά ερωτήματα όπως για παράδειγμα είναι τα εξής:

- Ποιοι δρόμοι έχουν κυκλοφοριακό πρόβλημα και ποια είναι η εναλλακτική διαδρομή με βάση το σημείο που βρίσκονται?
- Ποια είναι τα ενδιαφέροντα ενός ανθρώπου που κάθεται σε ένα υπολογιστή? Ποια sites κινούν το ενδιαφέρον του και πώς το περιεχόμενο αυτών μπορεί να γίνει πιο ελκυστικό?

- Δείχνουν τα ιατρικά δεδομένα εξάπλωση κάποιας νέας νόσου σε κάποιο σημείο του πλανήτη?
- Ποιο είναι το σημείο συνάντησης των φίλων ενός ατόμου?
- Ποια είναι τα πιο δημοφιλή θέματα στο διαδίκτυο (π.χ. Facebook, Twitter) και από ποιόν κατευθύνονται αυτές οι συζητήσεις?
- Πως επηρεάζονται τα χρηματιστήρια των χωρών μεταξύ τους κατά την διάρκεια της ημέρας?

Αυτά είναι μόνο ορισμένα ερωτήματα που μπορούν να απαντηθούν από το μεγάλο πλήθος εφαρμογών κινητού υπολογισμού [7].

Η πρόβλεψη των προτιμήσεων των χρηστών και η παροχή εξατομικευμένων υπηρεσιών ή προϊόντων βάσει αυτών των προτιμήσεων στηρίζεται στην ανάπτυξη εφαρμογών που βασίζονται στην πληροφορία - πλαισίου (context-aware computing). Ως πληροφορία πλαισίου ορίζεται «οποιαδήποτε πληροφορία μπορεί να περιγράψει την κατάσταση μιας οντότητας. Οντότητα μπορεί να αποτελεί μια συσκευή, μια τοποθεσία, ή ένας άνθρωπος που συσχετίζεται με την διάδραση ανάμεσα στον χρήστη και την εφαρμογή». Η επίγνωση πληροφορίας πλαισίου έχει να κάνει με την ικανότητα ενός συστήματος να κατανοεί αυτό που προσπαθεί να επιτύχει ο χρήστης και να προβαίνει στις κατάλληλες ενέργειες ώστε να του παρέχει τις υπηρεσίες/πληροφορίες μπορεί να φανούν χρήσιμες [8].

1.2.3 Απαιτήσεις

Πριν από μερικά χρόνια, η χρήση των εφαρμογών κινητού υπολογισμού ήταν αδιανόητο να επιτευχθεί. Ο βασικότερος λόγος ήταν η έλλειψη δεδομένων. Στις μέρες μας, οι εφαρμογές κινητού υπολογισμού χρησιμοποιούν δεδομένα που έχουν ψηφιακή μορφή και συλλέγονται κυρίως από αισθητήρες. Ωστόσο υπάρχουν αρκετά προβλήματα που προκύπτουν όταν γίνεται χρήση μεγάλων ροών δεδομένων όπως για παράδειγμα η αντιμετώπιση του «θορύβου» στα δεδομένα, η διόρθωση των λαθών και οι υπολογιστικές απαιτήσεις που χρειάζονται για την επεξεργασία των δεδομένων [7]. Οι πιο σημαντικές απαιτήσεις των εφαρμογών κινητού υπολογισμού είναι οι ακόλουθες:

- Ομοιογένεια στην μορφή των δεδομένων που συλλέγονται από τα διαφορετικά δίκτυα αισθητήρων.
- Άμεση καταγραφή πληροφορίας που συλλέγεται. Οι ροές δεδομένων μπορούν να ελεγχθούν μόνο κατά το διάστημα της συλλογής τους. Αν δεν γίνει εγκαίρως σύνοψη/ συγκέντρωση των δεδομένων, δεν θα είναι δυνατός ο ανασχηματισμός της πληροφορίας.
- Κλίμακα Δεδομένων: Η κλίμακα των δεδομένων είναι ένα ζήτημα για τις ροές δεδομένων που έχουν ως είσοδο τεράστιο όγκο και πρέπει να συνδεθούν με μεγάλες στατικές βάσεις δεδομένων. Σε πολλές εφαρμογές, είναι επαρκές ένα περιορισμένο πλήθος δεδομένων για την ολοκλήρωση μιας διεργασίας. Σε αυτές τις περιπτώσεις, πρέπει να λαμβάνονται δείγματα από τα δεδομένα κατά προσέγγιση.

- Συνεχής ενημέρωση (real- time constraints). Τα δεδομένα που συλλέγονται πρέπει να επεξεργάζονται συνεχώς και σε πραγματικό χρόνο ώστε να δίνουν απαντήσεις στο χρήστη πριν η πληροφορία απολέσει τη χρονική της εγκυρότητα.
- Συνεχή επεξεργασία του όγκου δεδομένων. Οι ροές δεδομένων απαιτούν συνεχή επεξεργασία γιατί συνήθως τα ερωτήματα που πρέπει να απαντηθούν (queries) είναι αποθηκευμένα και παραμένουν συνεχώς ενεργά ενώ τα δεδομένα έχουν συνεχή εισροή στην εφαρμογή.

1.2.4 Περιορισμοί

Ο σχεδιασμός ενός δικτύου αισθητήρων (όπως περιγράφτηκε στην ενότητα 1.2.1) επηρεάζεται από πολλούς παράγοντες όπως είναι η κατανάλωση ενέργειας, οι περιορισμοί του υλικού και του μέσου μετάδοσης, το λειτουργικό σύστημα, το κόστος παραγωγής, η δυνατότητα κλιμάκωσης, η τοπολογία του δικτύου κλπ. Οι σημαντικότεροι περιορισμοί περιγράφονται αναλυτικότερα παρακάτω.

Κατανάλωση Ενέργειας

Οι περιορισμοί ενέργειας σε ένα ασύρματο δίκτυο αισθητήρων είναι πολύ πιο αυστηροί σε σχέση με οποιοδήποτε άλλο ασύρματο δίκτυο. Ο λόγος είναι πως οι αισθητήρες οφείλουν να λειτουργούν σε ακραίες περιβαλλοντικές και γεωγραφικές συνθήκες, με την ελάχιστη ή ακόμα και καθόλου ανθρώπινη επίβλεψη και συντήρηση. Σε ορισμένες περιπτώσεις, η επαναφόρτιση της πηγής ενέργειας είναι αδύνατη. Η λειτουργία δικτύων με περιορισμένη ενέργεια, δηλαδή με αισθητήρες που τροφοδοτούνται από μπαταρίες, απαιτεί την χρήση ενός αποτελεσματικού πρωτόκολλου δικτύου, ζεύξης και φυσικού επιπέδου. Οι πηγές ενέργειας που χρησιμοποιούνται στα ασύρματα δίκτυα αισθητήρων μπορούν να κατηγοριοποιηθούν σε επαναφορτιζόμενες, μη-επαναφορτιζόμενες και ανανεώσιμες (δηλαδή χρησιμοποιούν τις παραμέτρους που είναι υπό μέτρηση για να ανανεώνουν την ενέργεια τους) [3].

Δυνατότητα Κλιμάκωσης

Ο αριθμός των κόμβων των αισθητήρων που χρησιμοποιούνται για τη μελέτη ενός φαινομένου μπορεί να είναι της τάξης των εκατοντάδων ή χιλιάδων και ανάλογα με την εφαρμογή, ο αριθμός να φτάσει και σε μια ακραία τιμή των εκατομμυρίων. Ο σχεδιασμός του δικτύου πρέπει να είναι σε θέση να ανταπεξέλθει σε αυτόν τον αριθμό των κόμβων και να αξιοποιήσει την υψηλή πυκνότητα του δικτύου αισθητήρων [6].

Υπολογιστική Ισχύς

Ένας αισθητήρας αποτελείται από τέσσερα βασικά στοιχεία: την ανιχνευτική μονάδα, την μονάδα επεξεργασίας, τον πομποδέκτη (που συνδέει τον αισθητήρα με το υπόλοιπο δίκτυο) και την μονάδα ισχύος. Συνήθως, περιλαμβάνουν και πρόσθετα στοιχεία ανάλογα με την εφαρμογή, π.χ. σύστημα ανίχνευσης θέσης, γεννήτρια,

κινητήρα (σε περίπτωση που χρειάζεται μετακίνηση του αισθητήρα για την συλλογή δεδομένων). Η μονάδα επεξεργασίας περιλαμβάνει μια μικρή αποθηκευτική μονάδα η οποία διαχειρίζεται τις διαδικασίες που επιτρέπουν στον κάθε ένα αισθητήρα να συνεργάζεται με τους υπόλοιπους και να εκτελεί τα καθήκοντα που του έχουν ανατεθεί. Όλα αυτά τα στοιχεία του αισθητήρα πρέπει να χωρέσουν σε μια μονάδα πολύ μικρού μεγέθους. Εκτός από το μέγεθος, υπάρχουν και άλλοι αυστηροί περιορισμοί για τους κόμβους των αισθητήρων. Πρέπει να καταναλώνουν πολύ χαμηλή ισχύ, να λειτουργούν σε υψηλές πυκνότητες όγκου, να έχουν χαμηλό παραγωγικό κόστος, να είναι αυτόνομοι και να προσαρμόζονται στο περιβάλλον [6].

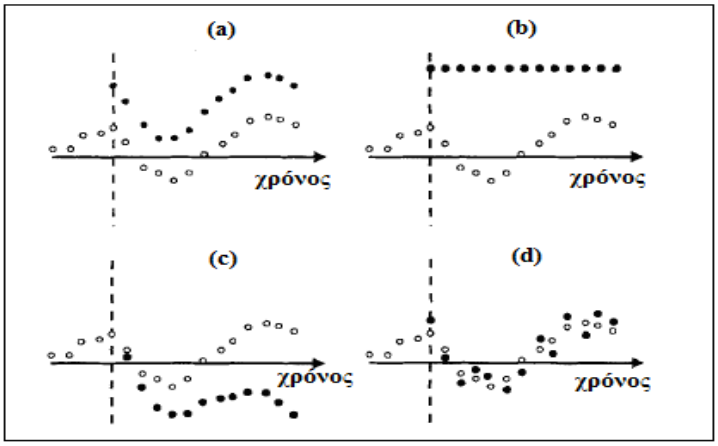
1.2.5 Σφάλματα αισθητήρων

Οι πληροφορίες που συλλέγονται από τα δίκτυα αισθητήρων χρησιμοποιούνται από πολλές εφαρμογές κινητού υπολογισμού. Για να εξαχθούν συμπεράσματα που θα είναι χρήσιμα στους χρήστες, εφαρμόζονται μέθοδοι εξόρυξης γνώσης στις ροές δεδομένων. Ωστόσο υπάρχουν ορισμένα προβλήματα στην εφαρμογή αυτών των μεθόδων. Τα προβλήματα που προκύπτουν προέρχονται κυρίως από περιττά δεδομένα που συλλέγονται από αισθητήρες (noisy data), από ελλιπή δεδομένα (incomplete data) και από την ανομοιογένεια που πιθανώς υπάρχει στη σημασιολογική ερμηνεία δεδομένων από την πλευρά των δικτύων.

Τα παραπάνω προβλήματα οφείλονται σε σφάλματα αισθητήρων. Υπάρχει περίπτωση οι αισθητήρες να έχουν καταστραφεί ή λόγω φυσικής καταπόνησης να μη μεταδίδουν σωστά δεδομένα. Τα σφάλματα γίνονται αντιληπτά είτε με μια μέτρηση που αποκλίνει από τις συνηθισμένες τιμές (outlier) είτε με αδυναμία αποστολής τιμής.

Πιο συγκεκριμένα, τα σφάλματα των αισθητήρων, μπορούν να χωριστούν στις εξής κατηγορίες (Σχ.1.2) [9]:

a) Μεροληψία στα συλλεχθέντα δεδομένα: Στις πραγματικές τιμές έχει προστεθεί μια σταθερά και η συνολική ένδειξη σφάλει.
b) Πλήρη αποτυχία μετάδοσης της τιμής. Ελλειπούσα τιμή ή μετάδοση μιας σταθεράς.
c) Συσσωρευτική απόκλιση σφαλμάτων. Σε κάθε ένδειξη προστίθεται το σφάλμα όλων των προηγούμενων μετρήσεων.
d) Υποβάθμιση της ακρίβειας της μέτρησης.



Εικόνα 2: Κατηγορίες Σφαλμάτων Αισθητήρων

Τα σφάλματα μπορούν να γίνουν άμεσα αντιληπτά, όπως π.χ. μία ελλείπουσα τιμή, ή χρονική καθυστέρηση στην μετάδοση, ενώ κάποια άλλα σφάλματα μπορεί να μη γίνουν αντιληπτά όπως π.χ. η υποβάθμιση της ακρίβειας και έτσι να επηρεάσουν αρνητικά την παρακολούθηση και τον έλεγχο της διαδικασίας.

Στα πλαίσια της παρούσας εργασίας θα μελετηθούν περιπτώσεις μετάδοσης μεμονωμένων τιμών αλλά και περιπτώσεις πλήρους αποτυχίας μετάδοσης τιμών.

1.3 Κίνητρο/Σκοπός της Εργασίας

Σκοπός της διπλωματικής εργασίας είναι να γίνει μια μελέτη των μηχανισμών εκτίμησης ελλιπούς πληροφορίας από δεδομένα που συλλέγονται από ασύρματα δίκτυα αισθητήρων. Έπειτα, να υπάρξει μια σύγκριση των αποτελεσμάτων αυτών των μεθόδων και να διεξαχθούν συμπεράσματα σχετικά με τα ποσοστά επιτυχίας της κάθε μεθόδου. Πιο αναλυτικά:

Στο πρώτο κεφάλαιο γίνεται μια εισαγωγή στην έννοια του διάχυτου υπολογισμού. Βλέπουμε την εξέλιξη της σχέσης μεταξύ ανθρώπου – υπολογιστή και με ποιους τρόπους η τεχνολογία και οι εφαρμογές της έχουν εισχωρήσει στην καθημερινότητα των ανθρώπων. Έπειτα, γίνεται μια περιγραφή της λειτουργίας των ασύρματων δικτύων αισθητήρων και της χρησιμότητάς τους στις εφαρμογές κινητού υπολογισμού. Παρουσιάζονται οι απαιτήσεις και οι περιορισμοί στη χρήση των αισθητήρων καθώς και τα σφάλματα που πρέπει να αντιμετωπιστούν για να μπορέσουμε να διεξάγουμε ορθά συμπεράσματα.

Στο δεύτερο κεφάλαιο γίνεται μια περιγραφή μεθοδολογιών διαχείρισης πληροφορίας για εφαρμογές κινητού υπολογισμού. Ο όγκος των δεδομένων που συλλέγονται από αισθητήρες είναι τόσο μεγάλος που είναι απαραίτητο να εφαρμοστούν τεχνικές μείωσης (ελάττωσης) των δεδομένων. Η σημαντικότερη τεχνική που χρησιμοποιείται είναι η ανάλυση κύριων συνιστωσών (PCA – Principal Component Analysis) όπου πραγματοποιείται ένας ορθογώνιος μετασχηματισμός των δεδομένων με σκοπό να επιλεχθούν ορισμένες κύριες συνιστώσες που μπορούν να απεικονίσουν σε ένα μεγάλο ποσοστό το αρχικό δείγμα βάσει των συσχετισμών μεταβολής του. Θα μελετηθούν αλγόριθμοι και εργαλεία που υλοποιούν την ανάλυση κύριων συνιστωσών.

Στο τρίτο κεφάλαιο εξετάζονται μεθοδολογίες εκτίμησης ελλιπούς πληροφορίας από αισθητήρες που δε μεταδίδουν καθόλου τιμές, δηλαδή δεν υπάρχει καμία

πρόβλεψη τιμής. Οι σημαντικότερες μεθοδολογίες είναι της παρεμβολής (interpolation), όπου η πρόβλεψη μιας τιμής καθορίζεται μέσα από ένα ήδη γνωστό εύρος τιμών των υπολοίπων παραμέτρων και της προεκβολής (extrapolation), όπου οι τιμές των παραμέτρων που καθορίζουν την ελλιπή τιμή βρίσκονται εκτός των αρχικών μετρήσεων.

Στο τέταρτο κεφάλαιο εξετάζονται ορισμένοι αλγόριθμοι πρόβλεψης μεμονωμένων ελλιπών τιμών από ένα data set. Οι αλγόριθμοι που παρουσιάζονται είναι οι εξής: C4.5 (υπο-περίπτωση του αλγορίθμου Decision Tree id3 που παίρνει ως είσοδο διακριτές τιμές), M5P, RepTree και Decision Stump. Επίσης, παρουσιάζονται τα εργαλεία που χρησιμοποιήθηκαν για την εφαρμογή αυτών των αλγορίθμων.

Στο πέμπτο κεφάλαιο παρουσιάζονται τα πειράματα που έγιναν έπειτα από την εφαρμογή των αλγορίθμων που μελετήθηκαν στα δύο προηγούμενα κεφάλαια. Περιγράφονται αναλυτικά τα δοκιμαστικά δεδομένα που χρησιμοποιήθηκαν, τα σενάρια που δοκιμάστηκαν και τα αποτελέσματα που διεξήχθησαν.

Στο έκτο κεφάλαιο γίνεται μια συγκεντρωτική περιγραφή όλων των συμπερασμάτων από την εφαρμογή των μεθοδολογιών που μελετήθηκαν και συνοψίζεται η ερευνητική συνεισφορά της παρούσας διπλωματικής εργασίας. Τέλος, παρουσιάζονται θέματα και ενότητες που επιδέχονται περαιτέρω μελέτης.

2 ΔΙΑΧΕΙΡΙΣΗ ΠΛΗΡΟΦΟΡΙΑΣ ΣΕ ΠΕΡΙΒΑΛΛΟΝΤΑ ΚΙΝΗΤΟΥ ΥΠΟΛΟΓΙΣΜΟΥ

2.1 Τεχνικές Μείωσης Διαστάσεων

Όπως αναφέρθηκε και στο προηγούμενο κεφάλαιο, ένα ασύρματο δίκτυο «έξυπνων» αισθητήρων επεξεργάζεται τα δεδομένα που συλλέγει. Δεδομένου ότι η υπολογιστική ισχύ των αισθητήρων είναι περιορισμένη και ο όγκος των δεδομένων τεράστιος, επιβάλλεται να εφαρμοστούν τεχνικές μείωσης διαστάσεων των αρχικών δεδομένων. Η μείωση των διαστάσεων είναι η διαδικασία κατά την οποία μειώνονται οι αρχικές μεταβλητές ενός συνόλου δεδομένων. Αυτή η διαδικασία μπορεί να γίνει είτε με την επιλογή γνωρισμάτων (feature selection) είτε με την εξαγωγή γνωρισμάτων (feature extraction). Οι μέθοδοι επιλογής γνωρισμάτων προσπαθούν να βρουν ένα υποσύνολο από τις αρχικές μεταβλητές που να μπορεί να αντιπροσωπεύσει το σύνολο των δεδομένων. Η εξαγωγή γνωρισμάτων προσπαθεί να προβάλει ένα σύνολο από διανύσματα υψηλής διάστασης σε ένα χώρο χαμηλότερης διάστασης. Σε αυτό το κεφάλαιο θα ασχοληθούμε με μεθοδολογίες εξαγωγής γνωρισμάτων.

2.1.1 Βιβλιογραφική ανασκόπηση μεθόδων και αλγορίθμων

Τα τελευταία χρόνια, υπάρχει μεγάλη ανάπτυξη στις μεθοδολογίες συλλογής δεδομένων και στα μέσα αποθήκευσής τους. Η δυνατότητα που παρέχεται σε πολλούς τομείς επιστημών, να συλλέγουν όλο και περισσότερες πληροφορίες για να διεξάγουν τα συμπεράσματά τους, οδηγεί σε μια υπερφόρτωση από πληροφορίες προς επεξεργασία. Οι ερευνητές που εργάζονται σε τομείς, όπως η μηχανική, η βιολογία, η οικονομία κ.α., έχουν να αντιμετωπίσουν όλο και μεγαλύτερο όγκο δεδομένων και παρατηρήσεων σε καθημερινή βάση. Αυτές οι νέες βάσεις δεδομένων θέτουν νέες προκλήσεις στην ανάλυση των δεδομένων. Τα παραδοσιακά στατιστικά μοντέλα δεν μπορούν να ανταπεξέλθουν στις νέες απαιτήσεις λόγω της αύξησης του αριθμού των παρατηρήσεων, αλλά κυρίως λόγω της αύξησης του αριθμού των μεταβλητών που σχετίζονται με κάθε παρατήρηση. Οι διαστάσεις των δεδομένων είναι ο αριθμός των μεταβλητών που μετρώνται σε κάθε παρατήρηση.

Ένα από τα σημαντικότερα προβλήματα των πολλαπλών διαστάσεων συνόλων δεδομένων είναι ότι σε πολλές περιπτώσεις, κάποιες από τις μεταβλητές δεν είναι απαραίτητες για την κατανόηση των φαινομένων που παρουσιάζουν ενδιαφέρον. Αν και ορισμένες υπολογιστικές μέθοδοι μπορούν να κατασκευάσουν μοντέλα πρόβλεψης με μεγάλη ακρίβεια από πολλαπλών διαστάσεων δεδομένα, εξακολουθεί να αποτελεί ενδιαφέρον σε πολλές εφαρμογές η μείωση των διαστάσεων των αρχικών μεταβλητών πριν από οποιαδήποτε μοντελοποίηση των δεδομένων [18].

Όταν τα δεδομένα εισόδου σε έναν αλγόριθμο είναι πολύ μεγάλα και υπάρχει η υποψία ότι πολλά από αυτά τα δεδομένα είναι περιττά (δηλαδή τα περισσότερα δεν παρέχουν νέα πληροφορία), για να μπορέσει να γίνει η επεξεργασία τους πρέπει να μετασχηματιστούν σε μια νέα, μειωμένη σειρά από χαρακτηριστικά γνωρίσματα (που ονομάζονται feature vectors). Ο μετασχηματισμός των δεδομένων εισόδου σε ένα σύνολο χαρακτηριστικών γνωρισμάτων ονομάζεται feature extraction. Εάν επιλεχθεί προσεκτικά το πλήθος των εξαγόμενων χαρακτηριστικών γνωρισμάτων, αναμένεται ότι το νέο σύνολο μπορεί να εξάγει όλες τις απαραίτητες πληροφορίες από τα δεδομένα

εισόδου προκειμένου να εκτελέσει την επιθυμητή εργασία έναντι του αρχικού μεγέθους δεδομένων εισόδου.

Οι τεχνικές μείωσης διαστάσεων παρέχουν λύση στο πρόβλημα διαχείρισης δεδομένων πολλαπλών διαστάσεων, αναζητώντας μια δομή χαμηλότερης διάστασης στα πολυδιάστατα δεδομένα. Ο μετασχηματισμός των δεδομένων μπορεί να είναι γραμμικός, π.χ. Ανάλυση Κύριων Συνιστωσών, Common Factor Analysis αλλά υπάρχουν και μη – γραμμικές τεχνικές μείωσης δεδομένων.

2.1.2 Κατηγορίες μεθόδων

Όπως αναφέρθηκε και στην προηγούμενη ενότητα οι τεχνικές μείωσης διαστάσεων κατηγοριοποιούνται σε γραμμικές και μη-γραμμικές. Στην παρούσα ενότητα θα εξετάσουμε ορισμένες από τις διαδεδομένες γραμμικές τεχνικές μείωσης διαστάσεων:

Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis - PCA)

Η ανάλυση σε κύριες συνιστώσες είναι μια τεχνική μείωσης του δείγματος. Χρησιμοποιείται όταν έχουμε ψηλά συσχετισμένες μεταβλητές. Μειώνει τον αριθμό των αρχικών μεταβλητών σε ένα μικρότερο αριθμό κύριων συνιστωσών που μετρούν τη μεγαλύτερη δυνατή διασπορά του δείγματος. Είναι μια διαδικασία που εφαρμόζεται σε μεγάλα δείγματα. Στην ενότητα 2.2 θα εξετάσουμε αναλυτικά τον τρόπο που λειτουργεί και τους αλγορίθμους που την υλοποιούν.

Παραγοντική Ανάλυση (Factor Analysis)

Η παραγοντική ανάλυση είναι μια τεχνική μείωσης των μεταβλητών ενός συνόλου δεδομένων, η οποία αναγνωρίζει τον αριθμό των λανθάνουσων δομών και δημιουργεί μια δομή, ένα νέο σύνολο μεταβλητών, τους κοινούς παράγοντες που ερμηνεύουν το δείγμα. Η εφαρμογή της παραγοντικής ανάλυσης υποθέτει ότι οι μετρήσιμες μεταβλητές εξαρτώνται από κάποιους άγνωστους, μη-μετρήσιμους παράγοντες. Σκοπός της παραγοντικής ανάλυσης είναι να ανακαλύψει τέτοιου είδους σχέσεις και να εκτιμήσει τους παράγοντες εκείνους που έχουν επίδραση και αντανακλούν τις αρχικές μεταβλητές. Αυτοί οι παράγοντες μπορούν να χρησιμοποιηθούν για να μειωθούν οι διαστάσεις του αρχικού συνόλου δεδομένων σύμφωνα με το παραγοντικό μοντέλο.

Σύμφωνα με το παραγοντικό μοντέλο [20], έστω ότι έχουμε x_1, x_2, x_3, …, x_p παρατηρήσεις που έχουν γίνει πάνω σε p μεταβλητές. Κάθε μια παρατήρηση μπορεί να γραφτεί ως γραμμική συνάρτηση m βασικών παραγόντων (όπου m < p) f_1, f_2, f_3, …, f_m συν ενός ειδικού παράγοντα (ειδικό σφάλμα) που οφείλεται για την αναξιοπιστία των μετρήσεων. Δηλαδή:

$$x_j = a_{j1} f_1 + a_{j2} f_2 + \dots + a_{jm} f_m + e_j \qquad j = 1, 2, \dots, p$$

όπου τα a_{j1}, a_{j2}, a_{jm} ονομάζονται factor loadings και τα e_j specific factors (ειδικοί παράγοντες).

2.1.3 Χρησιμότητα και πότε ενδείκνυνται

Η χρήση των τεχνικών μείωσης διαστάσεων είναι απαραίτητη γιατί το μέγεθος του συνόλου δεδομένων μειώνεται και έχει ως αποτέλεσμα τη γρηγορότερη επεξεργασία τους στον ελαττωμένο χώρο. Πολλές υπολογιστικές συσκευές έχουν περιορισμένες δυνατότητες και δε διαθέτουν μεγάλη υπολογιστική ισχύ, όπως είναι αυτές του κινητού υπολογισμού. Μόνο με την εφαρμογή τεχνικών μείωσης διαστάσεων θα μπορέσουν να αποθηκεύσουν την εισερχόμενη πληροφορία ώστε να χρησιμοποιηθεί κατάλληλα αργότερα. Επίσης, με την χρήση αυτών των τεχνικών αποκαλύπτεται η δομή των δεδομένων η οποία παραμένει κρυμμένη στον αρχικό πολυδιάστατο χώρο. Έτσι, βελτιώνεται η αποδοτικότητα των τεχνικών εξόρυξης γνώσης που εφαρμόζονται στα δεδομένα.

Πιο συγκεκριμένα η Ανάλυση Κύριων Συνιστωσών χρησιμοποιείται κυρίως για την μείωση των δεδομένων. Οι πρώτες κύριες συνιστώσες αρκούν για να αντικαταστήσουν την πλειονότητα των αρχικών δεδομένων. Η μείωση των διαστάσεων είναι επίσης χρήσιμη και όταν το πλήθος των παρατηρήσεων είναι μικρότερο από αυτό των μεταβλητών. Μια άλλη χρήση της ανάλυσης κύριων συνιστωσών είναι η οπτικοποίηση των δεδομένων. Η απεικόνιση των δεδομένων είναι δύσκολο να επιτευχθεί όταν υπάρχουν παραπάνω από τρείς μεταβλητές. Μέσω των κύριων συνιστωσών είναι εύκολο να απεικονιστεί η μεγαλύτερη μεταβλητότητα των δεδομένων σε δύο μόνο διαστάσεις. Η πολυδιάστατη πληροφορία μπορεί να συγχωνευτεί σε μια πιο συμπαγή απεικόνιση. Τέλος, η ανίχνευση ακραίων τιμών (noise reduction) και η ομαδοποίηση των δεδομένων γίνεται πιο εύκολη με την χρήση αυτής της τεχνικής [19] [20].

Παρότι η ανάλυση κύριων συνιστωσών και η παραγοντική ανάλυση δίνουν παρόμοια αποτελέσματα στην μείωση των διαστάσεων και την ερμηνεία των αρχικών δεδομένων διαφέρουν κατά πολύ στον τρόπο ανάλυσης των δεδομένων. Η ανάλυση κυριών συνιστωσών ερμηνεύει το μεγαλύτερο ποσοστό της ολικής διασποράς ενώ η παραγοντική ανάλυση εξυπηρετεί καλύτερα την ανάλυση του δείγματος διακρίνοντας ποσοτικές και ποιοτικές ομοιότητες. Στην ανάλυση κύριων συνιστωσών μας απασχολεί η συνολική διασπορά του συνόλου των δεδομένων ενώ στην παραγοντική ανάλυση το ενδιαφέρον επικεντρώνεται στο ποσοστό της ολικής διασποράς που μοιράζονται οι μεταβλητές. Κατά συνέπεια μια σημαντική διαφορά ανάμεσα στις δύο τεχνικές είναι το ποσοστό της συνολικής διασποράς που αναλύεται. Επίσης η παραγοντική ανάλυση μπορεί να ελέγξει εκ των προτέρων υποθέσεις για τον αριθμό των κοινών παραγόντων που απαιτούνται για την ερμηνεία του δείγματος.

2.2 Ανάλυση Κύριων Συνιστωσών (PCA)

Η Ανάλυση Κύριων Συνιστωσών είναι μια πολύ διαδεδομένη στατιστική τεχνική, που χρησιμοποιείται ευρέως σε προβλήματα που προκύπτουν από την επεξεργασία σήματος (π.χ. εξαγωγή βασικών χαρακτηριστικών, εκτίμηση ελλιπούς σήματος, ανίχνευση και διαχωρισμός στοιχείων του λόγου), αλλά και στην ανάλυση δεδομένων. Αυτή η μαθηματική διαδικασία χρησιμοποιεί τον ορθογώνιο μετασχηματισμό για να μετατρέψει ένα σύνολο από παρατηρήσεις, όπου το πλήθος των μεταβλητών του είναι πολύ μεγάλο και πιθανώς συσχετιζόμενο, σε ένα νέο σύνολο μη-συσχετισμένων

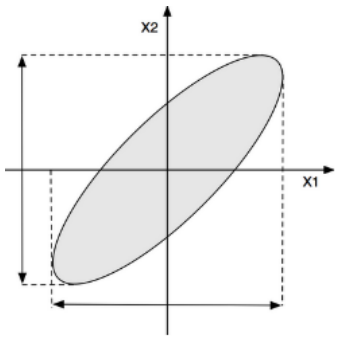
μεταβλητών, που ονομάζονται κύριες συνιστώσες (principal components). Από τις νέες συνιστώσες, μόνο ένα μέρος αυτών θα χρησιμοποιηθεί για την εξαγωγή συμπερασμάτων. Η μελέτη δύο ή τριών μη- συσχετιζόμενων μεταβλητών είναι ευκολότερη από τη μελέτη του συνόλου των αρχικών μεταβλητών. Ο μετασχηματισμός γίνεται με τέτοιο τρόπο ώστε η πρώτη κύρια συνιστώσα να προκύπτει από τα δεδομένα με την μεγαλύτερη διακύμανση και κάθε επόμενη συνιστώσα να έχει την αμέσως επόμενη μεγαλύτερη δυνατή διακύμανση, με τον περιορισμό ότι οι κύριες συνιστώσες είναι μεταξύ τους ορθογώνιες (δηλαδή παραμένουν ασυσχέτιστες μεταξύ τους). Οι κύριες συνιστώσες είναι ανεξάρτητες (δεν συσχετίζονται) εάν το σύνολο των δεδομένων ακολουθεί κανονική κατανομή.

Η μέθοδος αυτή αρχικά περιγράφτηκε το 1901 από τον Karl Pearson και αναπτύχθηκε περισσότερο αργότερα από τον Hotelling. Η πρακτική χρήση της μεθόδου ακολούθησε μετά την ευρεία διάδοση των Η/Υ, επειδή οι μαθηματικές πράξεις ήταν πολύ δύσκολο να πραγματοποιηθούν για περισσότερες από τέσσερις μεταβλητές.

2.2.1 Θεωρητική μελέτη και ερμηνεία

Πολλές μεταβλητές από ένα σύνολο δεδομένων έχουν συχνά παρόμοια πορεία (δηλαδή αυξομειώνονται αντίστοιχα). Αυτό οφείλεται στο γεγονός ότι παραπάνω από μια μεταβλητή μετρά χαρακτηριστικά του συστήματος που υποκινούνται από τα ίδια γνωρίσματα. Σε αυτές τις περιπτώσεις, πρέπει να είναι δυνατόν να εξαλειφθούν οι περιττές πληροφορίες, δημιουργώντας ένα νέο σύνολο μεταβλητών που θα μπορεί να εξάγει μόνο τα ουσιαστικά χαρακτηριστικά του συστήματος. Η Ανάλυση Κύριων Συνιστωσών (PCA) είναι μια μέθοδος που κάνει ακριβώς αυτό. Δημιουργεί ένα νέο σύνολο μεταβλητών (χώρο βάσης), τις κύριες συνιστώσες, όπου κάθε μια είναι γραμμικός συνδυασμός των αρχικών μεταβλητών. Πρόκειται ουσιαστικά για ένα ειδικό μετασχηματισμό των δεδομένων. Μετατρέπει αριθμητικά δεδομένα σε ένα νέο σύστημα συντεταγμένων, όπου δεν καταγράφονται περιττές πληροφορίες στις μεταβλητές.

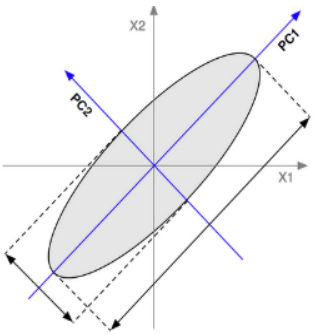
Στην Εικόνα 3 βλέπουμε μια τέτοια περίπτωση. Έστω ότι όλες οι τιμές του αρχικού δείγματος περικλείονται στο γραμμοσκιασμένο κομμάτι. Παρατηρούμε ότι οι δυο μεταβλητές που ορίζουν τα αρχικά δεδομένα (x_1 και x_2) συσχετίζονται μεταξύ τους. Όταν η μεταβλητή x_1 έχει μεγάλη τιμή τότε και η μεταβλητή x_2 έχει αντίστοιχα μεγάλη τιμή ενώ όταν η x_1 έχει μικρή τιμή τότε και η μεταβλητή x_2 έχει μικρή τιμή. Άρα, η διασπορά των δεδομένων είναι περίπου η ίδια και η κατανομή των δεδομένων ακολουθεί μια διαγώνια κλίση.



Εικόνα 3: Δισδιάστατη απεικόνιση των δεδομένων, χωρίς την χρήση PCA

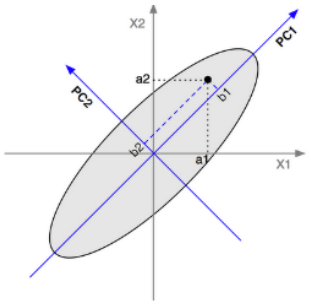
Η Ανάλυση Κύριων Συνιστωσών υπολογίζει ένα νέο σύνολο αξόνων (principal components – pc's) που περιγράφει τα δεδομένα πιο αποτελεσματικά. Η ιδέα είναι να περιστραφούν οι αρχικοί άξονες με τέτοιο τρόπο ώστε να καλύπτουν την μεγαλύτερη δυνατή διακύμανση των μεταβλητών. Ο πρώτος άξονας ονομάζεται πρώτη κύρια συνιστώσα (pc1) και έχει την κατεύθυνση της μεγαλύτερης διακύμανσης των δεδομένων. Κάθε νέος άξονας που δημιουργείται είναι ορθογώνιος με όλους τους προηγούμενους και έχει την κατεύθυνση της επόμενης μεγαλύτερης διακύμανσης.

Στην Εικόνα 4 βλέπουμε τις νέες κύριες συνιστώσες που έχουν δημιουργηθεί. Παρατηρούμε ότι η προβολή της pc1 είναι μεγαλύτερη από τις αρχικές μεταβλητές και αυτό οφείλεται στο γεγονός ότι καλύπτει μεγαλύτερη διασπορά στα δεδομένα. Επίσης η προβολή της pc1 είναι μεγαλύτερη από της pc2 γιατί είναι πιο σημαντική και μπορεί να περιγράψει περισσότερα δεδομένα. Αν υπήρχαν και άλλες κύριες συνιστώσες, η προβολή τους θα ήταν ολοένα και μικρότερη από των προηγούμενων συνιστωσών αφού θα αντιπροσώπευαν μικρότερη διασπορά στα αρχικά δεδομένα. Δηλαδή θα αντικατόπτριζαν όλο και πιο ασήμαντη πληροφορία. Με αυτό τον τρόπο μπορεί να γίνει συμπίεση των δεδομένων. Αγνοώντας τις λιγότερο σημαντικές κύριες συνιστώσες μπορούμε να έχουμε μια νέα εκδοχή των δεδομένων με λιγότερες μεταβλητές.



Εικόνα 4: Δημιουργία νέων Κύριων Συνιστωσών PC1 & PC2

Στην Εικόνα 5 παρουσιάζονται οι συντεταγμένες ενός σημείου στο αρχικό σύστημα αξόνων. Παρατηρούμε ότι με βάση τους άξονες των κύριων συνιστωσών μπορούν να διακριθούν περισσότερα σημεία απλά και μόνο κοιτάζοντας την pc_1, κάτι που δεν ισχύει στο αρχικό σύστημα συντεταγμένων. Δηλαδή υπάρχουν πολύ λιγότερα σημεία που ταυτοποιούνται με την τιμή b_1 απ'ότι με την a_1.



Εικόνα 5: Οι συντεταγμένες ενός σημείου στους άξονες PC.

Η χρήση των κύριων συνιστωσών μπορεί να χρησιμοποιηθεί σε διαφορετικού τύπου εφαρμογές. Οι πιο κοινές είναι: συμπίεση δεδομένων, γρήγορη αναζήτηση, οπτικοποίηση δεδομένων, εξεύρεση ομάδων και ακραίων τιμών.

Αλγεβρική Μελέτη

Παρακάτω ακολουθεί η αλγεβρική μελέτη για τον υπολογισμό των κύριων συνιστωσών όπως περιγράφεται στο [12].

Από την άλγεβρα γνωρίζουμε ότι ένας ορθογώνιος πίνακας U εάν πολλαπλασιαστεί με τον αντίστροφό του ισούται με τη μονάδα

$$U' U = 1 \qquad (1)$$

Άρα, ένας p × p συμμετρικός, μη-γραμμικός πίνακας, όπως ο πίνακας συνδιακύμανσης S, μπορεί να μειωθεί σε ένα διαγώνιο πίνακα L αν τον πολλαπλασιάσουμε με έναν συγκεκριμένο ορθογώνιο πίνακα U και τον αντίστροφό του, δηλαδή:

$$U'SU = L \qquad (2)$$

Τα διαγώνια στοιχεία του L ($l_1,l_2,\ldots,l_p$) ονομάζονται *ιδιοτιμές* του S. Οι στήλες του U ($u_1,u_2,\ldots,u_p$) ονομάζονται *ιδιοδιανύσματα* του S. Οι ιδιοτιμές μπορούν να υπολογιστούν από τη λύση της εξίσωσης:

$$|S - l\, I| = 0 \qquad (3)$$

όπου I είναι ο μοναδιαίος πίνακας.

Αυτή η εξίσωση παράγει μια πολυωνυμική εξίσωση p βαθμού στο *l*, απ' όπου μπορούμε να υπολογίσουμε τις τιμές $l_1,l_2,\ldots,l_p$.

<u>Για τον υπολογισμό των κύριων συνιστωσών:</u>

Έστω ότι έχουμε ένα αρχικό δείγμα δεδομένων με p μεταβλητές (p στήλες). Το πρώτο βήμα είναι να υπολογίσουμε τον πίνακα συνδιακύμανσης (covariance matrix) S:

$$S = \begin{pmatrix} s_1^2 & s_{12} & \ldots & s_{1p} \\ s_{12} & s_2^2 & \ldots & S_{2p} \\ \vdots & \vdots & & \vdots \\ s_{1p} & S_{2p} & \ldots & s_p^2 \end{pmatrix} \qquad (4)$$

όπου

s_i^2 : η διακύμανση της i μεταβλητής, x_i

s_{ij} : η συνδιακύμανση ανάμεσα στις i και j μεταβλητές

Εάν οι συνδιακυμάνσεις είναι διαφορετικές του μηδενός, σημαίνει ότι υπάρχει γραμμική συσχέτιση ανάμεσα στις δύο μεταβλητές. Η ισχύς αυτής της σχέσης δίνεται από τον τύπο:

$$r_{ij} = s_{ij} / (s_i\, s_j) \qquad (5)$$

Ο μετασχηματισμός των κύριων αξόνων θα μετατρέψει τις p συσχετισμένες μεταβλητές $x_1,x_2,\ldots,x_p$ σε p νέες, μη-συσχετιζόμενες μεταβλητές $z_1,z_2,\ldots,z_p$. Οι ισότιμοι άξονες των νέων μεταβλητών ορίζονται από τα χαρακτηριστικά διανύσματα u_1 που συνθέτουν την κατεύθυνση του συνημίτονου του πίνακα U που χρησιμοποιείται για το μετασχηματισμό:

$$Z = U' [\, x - \tilde{x} \,] \qquad (6)$$

όπου x και $\tilde{x}$ είναι p × 1 διανύσματα των παρατηρήσεων των αρχικών μεταβλητών και η μέση τιμή τους αντίστοιχα.

Επομένως, η i-οστή κύρια συνιστώσα, δίνεται από τον τύπο:

$$z_i = u'_i [x - \tilde{x}] \qquad (7)$$

Η μέση τιμή της i κύριας συνιστώσας είναι μηδέν και η διακύμανση l_i ίση με την i ιδιοτιμή.

2.2.2 Επιλογή των κύριων συνιστωσών

Μια από τις βασικότερες λειτουργίες της ανάλυσης κύριων συνιστωσών είναι η δυνατότητα να αναπαριστά ένα σύνολο δεδομένων p μεταβλητών σε k διαστάσεις, όπου k < p. Το ερώτημα που τίθεται είναι, ποιά είναι η βέλτιστη τιμή για το k. Προφανώς, όσο μεγαλύτερο είναι το k τόσο πιο αντιπροσωπευτικό και σύνθετο είναι το pca μοντέλο που δημιουργείται, ενώ όσο μικρότερο είναι το k τόσο πιο απλό θα είναι το μοντέλο. Για να καθοριστεί η τιμή του k υπάρχουν αρκετά κριτήρια. Τα κριτήρια αυτά ποικίλουν από σημαντικές στατιστικές μελέτες έως γραφικές διαδικασίες. Παρακάτω αναφέρονται ορισμένες από τις πιο διαδεδομένες μεθόδους [12],[13].

Μεθοδολογία Kaiser – Guttman

Ο πιο συνηθισμένος κανόνας τερματισμού στην ανάλυση κύριων συνιστωσών βασίζεται στην μέση τιμή των ιδιοτιμών. Επειδή συνήθως οι μεταβλητές μετριούνται σε διαφορετικές μονάδες μέτρησης, στην pca χρησιμοποιούνται οι πίνακες συσχέτισης έτσι ώστε να δίνεται το κατάλληλο βάρος σε κάθε μεταβλητή. Ως αποτέλεσμα, το πλήθος των ιδιοτιμών ισούται με το πλήθος των μεταβλητών. Σύμφωνα με την μεθοδολογία Kaiser – Guttman, διατηρούνται οι ιδιοτιμές που έχουν τιμή μεγαλύτερη από το μέσο όρο τους (π.χ. λ > 1), γιατί αυτοί οι άξονες συγκεντρώνουν περισσότερη πληροφορία από οποιαδήποτε άλλη αρχική μεταβλητή. Να σημειωθεί ότι αυτή η μεθοδολογία έχει επικριθεί αρκετά γιατί ένα μοντέλο pca που δημιουργείται από τυχαίες, ασυσχέτιστες μεταβλητές, παράγει ιδιοτιμές μεγαλύτερες από την μονάδα. Ωστόσο, εξακολουθεί να παραμένει μια από τις πιο δημοφιλείς μεθοδολογίες τερματισμού.

Scree Plot

Το scree plot είναι μια τεχνική γραφικής απεικόνισης, αρκετά διαδεδομένη που προτάθηκε από τον Cattell το 1966. Για την εφαρμογή αυτής της μεθόδου, σχεδιάζεται το γράφημα που προκύπτει από τις ιδιοτιμές και την τάξη της κάθε ιδιοτιμής. Όσο μικραίνουν οι ιδιοτιμές, τείνουν να βρίσκονται κατά μήκος μιας ευθείας γραμμής. Στα γραφήματα αυτά υπάρχει ένα σημείο (διάσπαση) στο οποίο η κλίση των αντίστοιχων τμημάτων αριστερά και δεξιά αυτού διαφέρουν αισθητά. Αυτό το σημείο καθορίζει ποια στοιχεία αξίζει να ερμηνευθούν περεταίρω (αποτελεί την τελευταία από τις pc's που συμμετέχουν στο μοντέλο) ή παραμένουν ασήμαντα.

Τα προβλήματα που προκύπτουν από τη χρήση αυτής της μεθοδολογίας είναι καταρχήν ότι βασίζεται σε μια γραφική αναπαράσταση και έπειτα ότι υπάρχει περίπτωση να μην ξεκάθαρος ο διαχωρισμός στην γραμμή που ορίζεται στο διάγραμμα. Αυτό μπορεί να οφείλεται είτε στο ότι η γραμμή έχει κλίση ομαλής καμπύλης είτε να υπάρχουν παραπάνω από μια διασπάσεις στην γραμμή. Στη δεύτερη περίπτωση συνηθίζεται να χρησιμοποιούνται ως κύριες συνιστώσες τα σημεία που ορίζονται από την πρώτη διάσπαση.

Μέθοδος Broken Stick

Η μέθοδος αυτή βασίζεται στο γεγονός ότι εάν διασπάσουμε ένα τμήμα μιας μονάδας μήκους σε p τμήματα, το αναμενόμενο μήκος g_k του k μακρύτερου τμήματος είναι:

$$g_k = \frac{1}{p}\sum_{i=k}^{p}\frac{1}{i}$$

Ένας τρόπος για να αποφασίσουμε αν το μέρος της διακύμανσης της k κύριας συνιστώσας είναι αρκετά μεγάλο για να συμπεριληφθεί στο μοντέλο είναι η σύγκρισή της με το g_k.

Ποσοστό συνολικής διακύμανσης

Ένα άλλο κριτήριο για την εκτίμηση του πλήθους των κύριων συνιστωσών είναι η συγκαταλογή όλων των συνιστωσών μέχρι ενός αυθαίρετου ποσοστού της συνολικής διακύμανσης. Αυτή η μέθοδος, συνήθως, περιλαμβάνει της συνιστώσες που αποτελούν το 95% της συνολικής διακύμανσης. Παρότι αρκετοί στατιστικολόγοι συνηγορούν υπέρ αυτής της μεθόδου, ο Jackson [12] τοποθετείται κατά αυτής χαρακτηρίζοντάς την αβάσιμη και αναξιόπιστη.

Cross Validation

Μια άλλη πρόταση για την εκτίμηση του βέλτιστου αριθμού pc's είναι η μεθοδολογία του cross-validation που προτάθηκε από τους Wold (1976,1978) και Eastment και Krzanowski (1982). Η τεχνική στηρίζεται στην τυχαία διαίρεση του αρχικού δείγματος σε g ομάδες, με n/g παρατηρήσεις η κάθε μια ομάδα. Η πρώτη ομάδα αφαιρείται από το δείγμα και εφαρμόζεται η μέθοδος PCA στο εναπομείναν. Τα διανύσματα που αποκτούνται από αυτό το μειωμένο δείγμα χρησιμοποιούνται για να βρεθούν οι pc's του διαγραμμένου δείγματος. Έπειτα, η ομάδα που έχει αφαιρεθεί από το δείγμα ξαναπροστίθεται και αφαιρείται η επόμενη ομάδα. Η ίδια διαδικασία επαναλαμβάνεται για όλες τις ομάδες.

Αυτή η μεθοδολογία συνιστάται όταν υπάρχει πρόθεση να κατασκευαστεί ένα μοντέλο PCA το οποίο θα χρησιμοποιηθεί για την αξιολόγηση μελλοντικών δεδομένων. Συμπεραίνουμε ότι το Cross Validation δεν μπορεί να χρησιμοποιηθεί σ' ένα αυτοματοποιημένο υπολογιστικό σύστημα παρακολούθησης όπου το μοντέλο πρέπει

να ανανεώνεται αναδρομικά γιατί τα παλαιότερα δεδομένα δεν αντιπροσωπεύουν την τρέχουσα κατάσταση της διεργασίας. Θα μπορούσε όμως να χρησιμοποιηθεί σε ένα αυτοματοποιημένο υπολογιστικό σύστημα παρακολούθησης όπου το μοντέλο ανανεώνεται με την προσέγγιση ενός κυλιόμενου παράθυρου.

2.2.3 Αλγόριθμοι υπολογισμού της μεθόδου PCA

Όπως διαπιστώσαμε και στις προηγούμενες ενότητες, η μέθοδος των κύριων συνιστωσών χρησιμοποιείται ευρέως ως τεχνική μείωσης διαστάσεων σε σύνολα δεδομένων μεγάλου όγκου. Για τη δημιουργία του pca μοντέλου υπάρχουν αρκετοί αλγόριθμοι. Ο κλάδος της Στατιστικής προσφέρει μια πληθώρα από αλγόριθμους για τον υπολογισμό των ιδιοτιμών και των ιδιοδιανυσμάτων. Οι περισσότεροι από αυτούς τους αλγόριθμους υλοποιούνται σε δύο φάσεις. Πρώτα, γίνεται μια προκαταρκτική μείωση από την αρχική μορφή σε μια δομημένη μορφή. Έπειτα με μια επαναληπτική διαδικασία έχουμε την τελική σύγκλιση. Οι αλγόριθμοι διαφοροποιούνται κυρίως στην δεύτερη φάση.

Ορισμένοι από τους αλγόριθμους υπολογισμού των κύριων συνιστωσών είναι οι εξής:

<u>Αλγόριθμος QR</u>

Η βασική ιδέα του αλγορίθμου QR [17] στηρίζεται στην ιδιότητα κάθε μη ιδιόμορφου πίνακα, όπως είναι ο πίνακας συνδιακύμανσης $S \in R^{m \times m}$, να μπορεί να γραφεί ως γινόμενο δύο πινάκων, ενός ορθογώνιου $Q \in R^{m \times m}$ και ενός άνω τριγωνικού $R \in R^{m \times m}$:

$$S = S1 = Q_1 \cdot R_1 \qquad (1)$$

Αν αντιστρέψουμε τους όρους της (1), έχουμε :

$$S_2 = R_1 \cdot Q_1 \qquad (2)$$

Αν επαναλάβουμε τα βήματα (1) και (2) πολλαπλασιάζοντας αντίστροφα τους όρους τότε θα έχουμε:

$$S_3 = Q_2 \cdot R_2$$

$$\ldots\ldots$$

$$S_{k-1} = Q_k \cdot R_k$$

$$S_k = R_k \cdot Q_k \qquad (3)$$

Αφού όμως ο Q_1 είναι ορθογώνιος (συνεπώς ισχύει ότι $Q_1 Q_1^T = I$), ο (1) μπορεί να γραφτεί και ως :

$$R_1 = Q_1^T \cdot S_1 \qquad (4)$$

επομένως ο (2) γίνεται:

$$S_2 = Q_1^T \cdot S_1 \cdot Q_1 \qquad (5)$$

Με την παραπάνω κυκλική διαδικασία προκύπτει ότι ο S_k γράφεται ως:

$$S_r = (Q_1 \cdot Q_2 \cdot \ldots \cdot Q_r)^T \cdot S \cdot (Q_1 \cdot Q_2 \cdot \ldots \cdot Q_r) \qquad (6)$$

Άρα ο S_r συγκλίνει και τείνει να γίνει διαγώνιος, με τα στοιχεία της διαγωνίου να είναι οι ιδιοτιμές του S ($r \rightarrow \infty$) και ο πίνακας Q τείνει στον πίνακα των ιδιοδιανυσμάτων του S.

Singular Value Decomposition (SVD)

Στην ανάλυση κύριων συνιστωσών, βρίσκουμε την κατεύθυνση των δεδομένων με την μεγαλύτερη διακύμανση (π.χ. τα ιδιοδιανύσματα που αντιστοιχούν στις μεγαλύτερες ιδιοτιμές των του πίνακα συνδιακύμανσης) και προβάλουμε τα αρχικά δεδομένα σε αυτές τις νέες διαστάσεις. Η λογική που εφαρμόζεται αυτό είναι ότι η πληροφορία δεύτερης τάξης βρίσκεται σε αυτές τις κατευθύνσεις. Αν χαρακτηρίσουμε τον πίνακα των ιδιοδιανυσμάτων που είναι ταξινομημένος βάσει των ιδιοτιμών ως V τότε μπορούμε να μετασχηματίσουμε τα αρχικά δεδομένα σε συνάρτηση του V. Τώρα τα ιδιοδιανύσματα ονομάζονται κύριες συνιστώσες. Αν επιλέξουμε μόνο τις πρώτες a γραμμές από αυτό το μετασχηματισμό τότε θα έχουμε καταφέρει να προβάλουμε τα δεδομένα από p σε a διαστάσεις. [16]

Η μέθοδος SVD στηρίζεται στο ακόλουθο θεώρημα της γραμμικής άλγεβρας: Ένας οποιοσδήποτε πίνακας X, μπορεί να γραφτεί ως το γινόμενο ενός (m × a) ορθογώνιου πίνακα U επί έναν (a × a) διαγώνιο πίνακα Λ με θετικές ή μηδενικές τιμές (singular values) και τον ανάστροφο (a × p) ορθογώνιο πίνακα V.

$$X = U \cdot \Lambda \cdot V^T$$

Με a αναπαριστάται το πλήθος των κύριων συνιστωσών. Αν το πλήθος των μετρήσεων (m γραμμές) είναι μεγαλύτερο από το πλήθος των αρχικών μεταβλητών (p στήλες), δηλαδή m > p τότε το a = p, διαφορετικά a = m – 1. Συνήθως εφαρμόζουμε αυτές τις μεθόδους όταν έχουμε μεγάλο όγκο δεδομένων οπότε το σύνηθες είναι να ισχύει m > p και συνεπώς a = p. Για τον διαγώνιο πίνακα Λ, προκύπτουν μοναδικές τιμές λ_j (για j = 1,2,…,a) στην κύρια διαγώνιο, με $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_a$ να αντιπροσωπεύουν τις κύριες συνιστώσες. Οι πρώτες pc's περιλαμβάνουν την πιο σχετική πληροφορία ενώ οι υπόλοιπες αντιπροσωπεύουν κυρίως θόρυβο και λιγότερη ουσιώδη πληροφορία [15].

Αλγόριθμος NIPALS

Ο αλγόριθμος NIPALS (**N**on-linear **I**terative **P**artial **L**east **S**quares) αναπτύχθηκε από τον H.Wold, αρχικά για τον υπολογισμό των πρώτων pc's στην μέθοδο PCA και έπειτα χρησιμοποιήθηκε στην στατιστική μεθοδολογία μερικής παλινδρόμησης ελαχίστων τετραγώνων (PLS- **P**artial **L**east **S**quare regression). Είναι ο πιο συνηθισμένος αλγόριθμος για τον υπολογισμό των κύριων συνιστωσών σε ένα σύνολο δεδομένων. Αν συγκρίνουμε τον αλγόριθμο NIPALS με τον αλγόριθμο SVD, ο NIPALS δίνει μεγαλύτερη ακρίβεια στα αριθμητικά αποτελέσματα αλλά απαιτεί μεγαλύτερη υπολογιστική πολυπλοκότητα.

Η μέθοδος αυτή στηρίζεται στον διαχωρισμό των αρχικών δεδομένων (πίνακας X) σε μια δομημένη μορφή (TP^T) και στον θόρυβο (E) [14].

Δηλαδή:

$$X = TP^T + E \qquad \text{(Δομημένη Μορφή + Θόρυβος)}$$

όπου:

T: Scores. Πίνακας που περιγράφει τη συσχέτιση μεταξύ των στηλών του X. Στον πίνακα T η πρώτη στήλη περιέχει τις τιμές της πρώτης κύριας συνιστώσας, η δεύτερη στήλη τις τιμές της δεύτερης κύριας συνιστώσας κ.ο.κ.

P: Loadings. Πίνακας με τα βάρη των μεταβλητών του X στον πίνακα scores T.Με την χρήση του πίνακα P, μπορούμε να δούμε ποιες μεταβλητές ευθύνονται για τα patterns που σχηματίζονται στον πίνακα T.

E: Noise. Πίνακας με το μέρος του θορύβου του PCA. Ο πίνακας E δεν αποτελεί μέρος του μοντέλου. Είναι το μέρος του πίνακα X που δεν μπορεί να αποτυπωθεί στο μοντέλο TP^T. Αν αυτός ο πίνακας είναι μεγάλος (κάτι που δεν θα πρέπει να ισχύει), σημαίνει ότι έχει αφαιρεθεί αρκετή πληροφορία από τα αρχικά δεδομένα κατά τον σχηματισμό του μοντέλου.

Ακολουθεί μια γενική επισκόπηση του αλγορίθμου NIPALS:

t: scores για την pc_i

p: loadings για την pc_i

threshold = 0,00001 (μια πολύ μικρή τιμή που θα χρησιμοποιηθεί για τον έλεγχο της σύγκλισης)

Θεωρούμε τον πίνακα $E_0 = X$.

Μέχρι να υπολογιστούν όλες οι PCs επαναλαμβάνουμε τα εξής βήματα:

1. Προβάλουμε τον X στον t με σκοπό να υπολογίσουμε την σχετική p
   $p = (E^T_{(i-1)}{}^t)=(t^T t)$

2. Κανονικοποιούμε το διάνυσμα p ώστε να έχει μήκος 1

   $p = p * (p^T p)^{-0.5}$

3. Βρίσκουμε την προβολή t του πίνακα X στο διάνυσμα p που υπολογίσαμε, με στόχο να βρούμε τα νέα scores

   $t = (E_{(i-1)}p) / (p^T p)$

4. Έλεγχος της σύγκλισης. Εάν η διαφορά των ιδιοτιμών $\tau_{new} = (t^T t)$ και τ_{old} (από την τελευταία επανάληψη) είναι μεγαλύτερη από την τιμή που μας δίνει το γινόμενο threshold * τ_{new}, τότε επαναλαμβάνουμε από το βήμα 1.

5. Αφαιρούμε την υπολογισμένη pc από τον πίνακα $E_{(i-1)}$

   $E_{(i)} = E_{(i-1)} - (tp^T)$

<u>Αλγόριθμος EM</u>

Ο αλγόριθμος E-M (Expectation - Maximization) είναι μια στατιστική μεθοδολογία που προσπαθεί να μεγιστοποιήσει τη πιθανότητα σωστής εκτίμησης μιας παραμέτρου (maximum-likelihood estimation) σε ένα στατιστικό μοντέλο όπου υπάρχουν μη - παρατηρούμενες μεταβλητές. Ο E-M είναι ένας επαναληπτικός αλγόριθμος που εναλλάσσεται μεταξύ δύο βημάτων, του e-step (expectation step) και m-step (maximization step). Το e-step υπολογίζει τις αναμενόμενες λογαριθμικές πιθανότητες (expectation log-likelihood), βάσει τις τρέχουσες εκτιμήσεις των παραμέτρων. Το m-step υπολογίζει τις παραμέτρους μεγιστοποιώντας τις αναμενόμενες λογαριθμικές πιθανότητες που υπολογίστηκαν στο e-step. Αυτές παράμετροι χρησιμοποιούνται στη συνέχεια για να προσδιοριστεί η κατανομή των μη-παρατηρούμενων μεταβλητών με στο επόμενο e-step.

Ο αλγόριθμος E-M μπορεί να χρησιμοποιηθεί για την εφαρμογή του PCA γιατί παρέχει έναν απλό και αποτελεσματικό τρόπο υπολογισμού λίγων ιδιοδιανυσμάτων και ιδιοτιμών όταν τα δεδομένα μετριούνται σε πολλές διαστάσεις. Επίσης, αυτός ο τρόπος υπολογισμού επιτρέπεται ακόμη και στη περίπτωση ελλιπών στοιχείων.

Άρα ο E-M αλγόριθμος για τον υπολογισμό των pc's προκύπτει από τα εξής βήματα:

e-step: $X = (C^T C)^{-1} C^T Y$

m-step: $C^{new} = Y X^T (X X^T)^{-1}$

όπου:

Y είναι ένας πίνακας (p x n) με τις τιμές των παραμέτρων και X ένας πίνακας (k x n) με τις άγνωστες τιμές. Οι στήλες του πίνακα C θα μας δώσουν τις k πρώτες pc's. Για τον αναλυτικό υπολογισμό των ιδιοτιμών και ιδιοδιανυσμάτων, προβάλουμε τα δεδομένα στο χώρο των k διαστάσεων και προκύπτει μια διατεταγμένη ορθογώνια βάση της συνδιακύμανσης [30].

2.3 Εργαλεία

Υπάρχουν αρκετές μέθοδοι μετασχηματισμού των βασικών παραμέτρων ενός συνόλου δεδομένων σε νέες μεταβλητές πρόβλεψης. Η χρησιμότητά τους έγκειται στην εύκολη και γρήγορη μείωση των διαστάσεων όταν υπάρχει δυσκολία στην διάταξη των αρχικών διαστάσεων. Σε αυτή την περίπτωση, ορισμένα χαρακτηριστικά, λιγότερο περιγραφικά, μπορούν να χρησιμοποιηθούν για την κατασκευή ενός νέου μοντέλου. Το Matlab είναι ένα εργαλείο που χρησιμοποιείται κατά κύριο λόγο για την επίλυση μαθηματικών προβλημάτων, ωστόσο είναι πολύ "ισχυρό" και μπορεί να χρησιμοποιηθεί και στον προγραμματισμό καθώς περιέχει αρκετές προγραμματιστικές εντολές και υλοποιημένες μεθόδους καθώς και πλατφόρμες διασύνδεσης με υπάρχουσες γλώσσες προγραμματισμού.

2.3.1 Matlab Statistic Toolbox

Το Matlab 7.11.0 (R2010b) είναι το εργαλείο που χρησιμοποιήθηκε στην παρούσα διπλωματική εργασία για τον υπολογισμό και την επιλογή των κύριων συνιστωσών (εφαρμογή pca) και πιο συγκεκριμένα η βιβλιοθήκη Matlab Statistic Toolbox [11]. Το Statistic Toolbox™ παρέχει ένα ολοκληρωμένο σύνολο εργαλείων για επεξεργασία και κατανόηση των δεδομένων. Επίσης, περιλαμβάνει μεθόδους και διαδραστικά εργαλεία για την μοντελοποίηση δεδομένων, ανάλυση ιστορικών τάσεων, προσομοίωση συστημάτων, ανάπτυξη στατιστικών αλγορίθμων και εκμάθηση / διδασκαλία της στατιστικής επιστήμης. Το Statistic Toolbox υποστηρίζει ένα ευρύ φάσμα λειτουργιών, από τον υπολογισμό βασικών μεταβλητών της στατιστικής έως την ανάπτυξη και την οπτικοποίηση πολυδιάστατων μη γραμμικών μοντέλων. Όλες οι λειτουργίες του Statistic Toolbox είναι γραμμένες σε ανοικτή γλώσσα MATLAB® ώστε να μπορεί να ελεγχθεί ο κάθε αλγόριθμος, να τροποποιηθεί ο πηγαίος κώδικας και να δημιουργηθούν νέες, προσαρμοσμένες συναρτήσεις.

Ας δούμε πιο συγκεκριμένα τις εντολές που χρησιμοποιήθηκαν για τον υπολογισμό των βασικών συνιστωσών:

Έστω ότι το αρχικό μας δείγμα είναι ο πίνακας δεδομένων ratings, με x πλήθος στηλών και y γραμμών. Οι μετρήσεις δεν έχουν υποστεί καμία επεξεργασία και έχουν γίνει με βάση διαφορετικής μονάδας μέτρησης.

Υπολογισμός των κύριων συνιστωσών:

```
[coefs,scores,variances,t2] = princomp(sr);
```

Από την χρήση της συνάρτησης princomp(…) προκύπτουν οι εξής τέσσερεις μεταβλητές:

coefs (Component Coefficients): Πίνακας που περιέχει τους συντελεστές του γραμμικού συνδυασμού των αρχικών μεταβλητών που παράγουν τις κύριες συνιστώσες.

scores (Component Scores): Περιέχει τις συντεταγμένες των αρχικών δεδομένων στο νέο σύστημα συντεταγμένων, όπως αυτό ορίζεται από τις κύριες συνιστώσες. Ο

πίνακας scores έχει το ίδιο μέγεθος με τα δεδομένα εισόδου. Από αυτό τον πίνακα επιλέγονται οι κύριες συνιστώσες που θα αντικαταστήσουν το αρχικό δείγμα δεδομένων.

variances (Component Variances): Διάνυσμα που περιέχει την διακύμανση της αντίστοιχης κύριας συνιστώσας. Η διακύμανση κάθε στήλης του πίνακα scores ισούται με το αντίστοιχο στοιχείο του πίνακα variances. Για να δούμε το ποσοστό της συνολικής διακύμανσης που αντιστοιχεί σε κάθε κύρια συνιστώσα χρησιμοποιούμε την εξής εντολή:

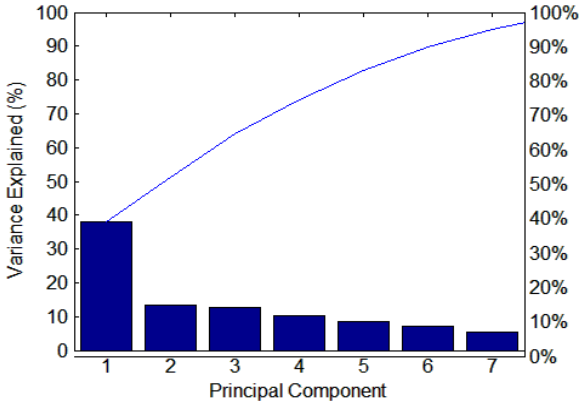
percent_explained = 100*variances/sum (variances)

Επίσης με την συνάρτηση pareto(…), μπορούμε να δούμε και γραφικά το ποσοστό της διακύμανσης κάθε κύριας συνιστώσας:

pareto(percent_explained)

xlabel('Principal Component')

ylabel('Variance Explained (%)')



Εικόνα 6: Γραφική απεικόνιση της συνάρτησης pareto()

t2 (Hotelling's T2): Ο πίνακας t2 είναι ένα στατιστικό μέτρο της πολυπαραγοντικής απόστασης κάθε μεταβλητής από το κέντρο του συνόλου δεδομένων. Αυτός είναι ένας τρόπος για να βρεθούν τα ακραία σημεία των δεδομένων.

3 ΕΚΤΙΜΗΣΗ ΕΛΛΙΠΟΥΣ ΠΛΗΡΟΦΟΡΙΑΣ ΑΠΟ ΕΛΑΤΤΩΜΑΤΙΚΟΥΣ ΑΙΣΘ/ΡΕΣ

3.1 Περιγραφή προβλήματος και εφαρμογές

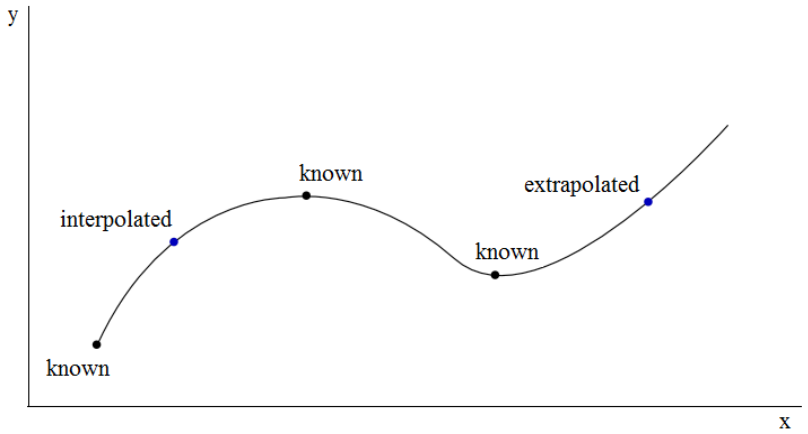
Όπως αναφέρθηκε στην ενότητα 1.2.5, πολλές φορές τα δεδομένα που συλλέγονται από αισθητήρες τείνουν να έχουν σφάλματα. Είναι σύνηθες φαινόμενο σε ένα δίκτυο αισθητήρων, κάποιος από τους αισθητήρες να παρουσιάσει πρόβλημα στην λειτουργία του και να αποτύχει να μεταδώσει την τιμή της μέτρησης. Άλλες φορές προκύπτουν προβλήματα στο ίδιο το δίκτυο με αποτέλεσμα να μην φθάνει η τιμή που σύλλεξε ο αισθητήρας στον σταθμό βάσης ή να φθάνει με μεγάλη καθυστέρηση. Όπως γνωρίζουμε, για την διεξαγωγή συμπερασμάτων πρέπει να εφαρμοστούν οι κατάλληλοι αλγόριθμοι. Το πρόβλημα παρουσιάζεται όταν τα δεδομένα που έχουν συλλεχθεί από ένα δίκτυο αισθητήρων για να χρησιμοποιηθούν ως είσοδο σε έναν αλγόριθμο είναι ελλιπή. Σε αυτό το κεφάλαιο θα μελετηθούν μεθοδολογίες εκτίμησης ελλιπούς πληροφορίας που προέρχονται από ελαττωματικούς αισθητήρες που μεταδίδουν ελλιπείς τιμές. Δηλαδή από αισθητήρες που μπορεί να βρεθούν εκτός λειτουργίας και συνεπώς να αποτύχουν να αποστείλουν τις τιμές μέτρησής τους.

3.2 Μεθοδολογίες Παρεμβολής / Προεκβολής (Interpolation /Extrapolation)

Κατά τη διάρκεια συλλογής των δεδομένων από τους αισθητήρες μπορούμε να δημιουργήσουμε μια καμπύλη πάνω στην οποία θα εμφανίζονται όλα τα συλλεχθέντα σημεία. ***Παρεμβολή (Interpolation)*** είναι η διαδικασία ανάκτηση τιμών από την καμπύλη ανάμεσα στα ήδη γνωστά σημεία (δεδομένα). Άρα, μια σειρά από σημεία δεδομένων, που έχουν ληφθεί από δειγματοληψία, αντιπροσωπεύουν τις τιμές μιας συνάρτησης για ένα πεπερασμένο αριθμό τιμών της ανεξάρτητης μεταβλητής. Συχνά απαιτείται η παρεμβολή (εκτίμηση) των τιμών της συνάρτησης για μία ενδιάμεση τιμή της ανεξάρτητης μεταβλητής. Αυτό επιτυγχάνεται μέσω της προσαρμογής της καμπύλης ή χρήση της παλινδρομικής ανάλυσης.

Προεκβολή (Extrapolation) είναι η διαδικασία απόκτησης μιας τιμής από ένα γράφημα ή μια γραφική παράσταση που εκτείνεται πέρα από τα συλλεχθέντα δεδομένα. Η διαδικασία της προεκβολής είναι παρόμοια με την διαδικασία της παρεμβολής αλλά τα αποτελέσματά της συνήθως υπόκεινται σε μεγαλύτερη αβεβαιότητα γιατί δεν περιορίζονται σε ένα γνωστό εύρος τιμών.

Στην Εικόνα 7 μπορούμε να δούμε διαγραμματικά την έκταση μιας καμπύλης που καλύπτει η μεθοδολογία της παρεμβολής και της προεκβολής.



Εικόνα 7: Παρεμβολή (Interpolation) / Προεκβολή (Extrapolation)

Συνοψίζοντας, παρεμβολή είναι η εκτίμηση μιας τιμής που βρίσκεται ανάμεσα σε δύο ήδη γνωστές τιμές μιας ακολουθίας τιμών, ενώ προεκβολή είναι η πρόβλεψη μιας τιμής που βασίζεται στην επέκταση της ακολουθίας των ήδη γνωστών τιμών ή γενικότερα μιας περιοχής τιμών που δεν είναι γνωστή.

3.2.1 Βασική Περιγραφή και ερμηνεία

Σε αρκετά πρακτικά προβλήματα που προκύπτουν, είναι σημαντικό να μπορεί να γίνει εκτίμηση όχι μόνο της τρέχουσας κατάστασης του συστήματος, αλλά και πρόβλεψη μιας μελλοντικής κατάστασης. Η πρόβλεψη αυτή πρέπει να βασίζεται στα αποτελέσματα των παρατηρήσεων της συμπεριφοράς του ίδιου του συστήματος. Οι προβλέψεις αυτές, είναι πολύ σημαντικές κυρίως για την επίλυση προβλημάτων ελέγχου αλλά και για την αποφυγή επιβλαβών καταστάσεων στη λειτουργία του συστήματος.

Το πρόβλημα λοιπόν της πρόβλεψης τιμών ορίζεται ως εξής:

Έστω ότι υπάρχει ένα σύνολο από n+1 πεπερασμένα σημεία (x_i, y_i) στο δισδιάστατο χώρο, όπου i = 0, … , n και για τα οποία οι συντεταγμένες είναι γνωστές. Το ζήτημα που τίθεται είναι πώς θα μπορέσουμε να υπολογίσουμε την τιμή y που αντιστοιχεί σε μία τιμή x, η οποία δεν ανήκει στο σύνολο των x_i, i = 0, … , n. Υπολογίζοντας μια συνάρτηση y = f(x) η οποία θα περνάει από όλα τα γνωστά σημεία (x_i, y_i), μπορούμε να εκτιμήσουμε (προβλέψουμε) την τιμή y που αντιστοιχεί στο x που δεν ανήκει στο σύνολο των x_i, i = 0, … , n

Η συνάρτηση που υπολογίσαμε, περνάει από όλα τα γνωστά σημεία, και συνεπώς ισχύει: $y_i = f(x_i)$, i = 0, … , n. Επίσης, προσεγγίζει τα σημεία τα οποία βρίσκονται ανάμεσα στα γνωστά σημεία. Έτσι, θα μπορούσαμε να πούμε ότι εκφράζει μία μαθηματική σχέση στο πρόβλημα που μελετάμε και έτσι μπορούμε να κάνουμε μια πρόβλεψη για τις τιμές του y, όταν το x πάρει τιμές που βρίσκονται στο διάστημα [x_0, x_n]. Αυτή η μεθοδολογία εκτίμησης τιμής ονομάζεται παρεμβολή (interpolation). Τα γνωστά σημεία είναι βέβαιο ότι επαληθεύονται από τη σχέση αυτή, αφού η συνάρτηση περνά από αυτά. Όμως, δεν σημαίνει ότι η τιμή της f(x) για μία ενδιάμεση τιμή x αντιπροσωπεύει την πραγματική τιμή y. Συνεπώς, η συνάρτηση παρεμβολής συνδέεται με σφάλματα ανάμεσα στις πραγματικές τιμές και στις τιμές που προσεγγίζει. Ο στόχος είναι να βρεθεί μια συνάρτηση παρεμβολής η οποία θα δίνει το μικρότερο δυνατό σφάλμα. Για να επιτευχτεί αυτός ο στόχος, έχουν αναπτυχθεί μεθοδολογίες

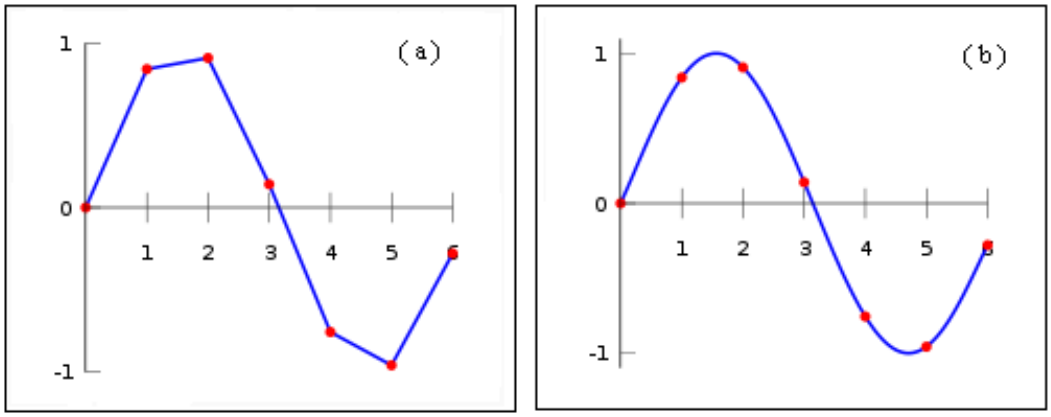
υπολογισμού της συνάρτησης f(x) που δεν περνάνε από όλα τα γνωστά σημεία. Έτσι προκύπτει μια πιο ομαλή συνάρτηση που περιγράφει καλύτερα τα δεδομένα. Συνήθως, τα αποτελέσματα που δίνουν αυτές οι μεθοδολογίες είναι ακριβέστερα.

Η προεκβολή (extrapolation) χρησιμοποιεί τις ίδιες συναρτήσεις που κατασκευάζονται για την παρεμβολή. Συνεπώς αυτές οι δύο μεθοδολογίες σχετίζονται άμεσα. Η διαφορά της προεκβολής με την παρεμβολή είναι ότι μια συνάρτηση y = f(x) χρησιμοποιείται για να εκτιμηθεί ένα μελλοντικό σημείο, δηλαδή ένα σημείο που δεν βρίσκεται στο εύρος που ορίζουν τα ήδη γνωστά σημεία (x_i, y_i). Ωστόσο η διαδικασία εύρεσης της συνάρτησης και η χρήση των γνωστών σημείων και των κατάλληλων μαθηματικών ακολουθεί την ίδια λογική. Όσο το x απομακρύνεται από το πεδίο τιμών των γνωστών x_i, είναι λογικό η προβλεπόμενη τιμή του y να εμπεριέχει μεγαλύτερο σφάλμα, αφού η συνάρτηση γίνεται πιο απρόβλεπτη.

Για την κατασκευή συναρτήσεων παρεμβολής / προεκβολής υπάρχουν αρκετές μεθοδολογίες. Η πιο απλή είναι η γραμμική συνάρτηση η οποία παίρνει δύο σημεία (x_a,y_a) και (x_b,y_b), και για να προβλέψει την τιμή του x στον y χρησιμοποιεί τον ακόλουθο τύπο:

$$y = y_a + (y_b - y_a)\frac{(X - X_a)}{(X_b - X_a)}$$

Η γραμμική συνάρτηση είναι γρήγορη και εύκολη αλλά δεν είναι αρκετά ακριβής. Η χρήση των πολυωνύμων στην γραμμική συνάρτηση μπορεί να δώσει πολύ καλύτερα αποτελέσματα και να ελαχιστοποιήσει τα σφάλματα. Στην Εικόνα 8 μπορούμε να δούμε και διαγραμματικά τις διαφορές μεταξύ πολυωνυμικών και μη-πολυωνυμικών συναρτήσεων.



Εικόνα 8: (a) Γραμμική Συνάρτηση **(b) Πολυωνυμική Συνάρτηση**

Οι πιο σημαντικοί πολυωνυμικοί αλγόριθμοι που χρησιμοποιούνται ευρέως για την πρόβλεψη τιμών καλύπτονται αναλυτικότερα στην επόμενη ενότητα.

3.2.2 Αλγόριθμοι

Για την εφαρμογή του Intrpolation / Extrapolation είναι αναπόφευκτη η χρήση των πολυωνύμων. Ωστόσο υπάρχουν διαφορετικές προσεγγίσεις και αλγόριθμοι που υλοποιούν την παρεμβολή και την προεκβολή. Η πιο συχνά χρησιμοποιούμενη μεθοδολογία είναι τα πολυώνυμα Lagrange που χρησιμοποιούνται σε περιπτώσεις μικρού πλήθους δεδομένων. Επίσης χρησιμοποιούνται για την προσέγγιση παραγώγων και ολοκληρωμάτων. Η μέθοδος του Newton χρησιμοποιείται κατά βάση στις υπολογιστικές μεθόδους των πολυωνύμων και στην επίλυση διαφορικών εξισώσεων. Ωστόσο, τα πολυώνυμα παρουσιάζουν αδυναμία λόγω της μεγάλης ταλάντωσης, ιδιαίτερα όταν το πλήθος τον μετρήσεων είναι πολύ μεγάλο.

Τα πολυώνυμα Hermite χρησιμοποιούνται για παρεμβολή σε συναρτήσεις και τις παραγώγους τους. Είναι αρκετά ακριβή αλλά απαιτούν περισσότερες πληροφορίες σχετικά με τη συνάρτηση που προσεγγίζουν. Όταν υπάρχει μεγάλος όγκος δεδομένων, προκύπτει επίσης το πρόβλημα της ταλάντωσης.

Η πιο συνηθισμένη μεθοδολογία στην εφαρμογή παρεμβολής και προεκβολής είναι η χρήση τμηματικών πολυωνύμων. Αν είναι γνωστή η συνάρτηση και οι παράγωγες τιμές της, τότε συνίσταται η χρήση της μεθόδου Cubic Hermite. Αν διατίθενται μόνο οι τιμές της συνάρτησης τότε χρησιμοποιείται η μέθοδος Cubic Spline.

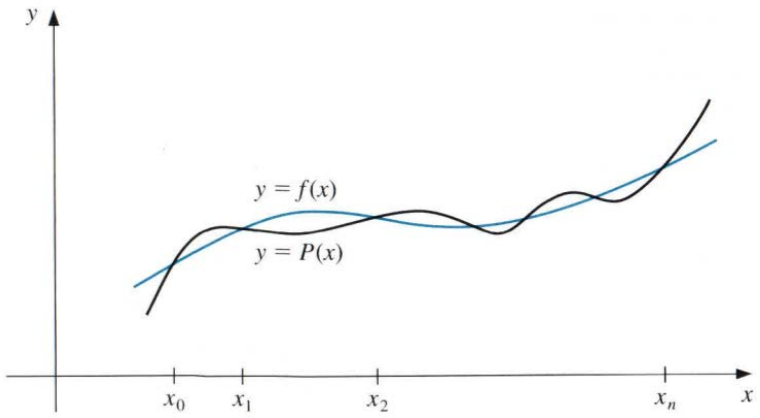
Συνήθως χρησιμοποιούνται και άλλες μέθοδοι, όπως αυτή της τριγωνομετρικής παρεμβολής. Αυτή η μέθοδος χρησιμοποιείται σε περιπτώσεις με μεγάλο πλήθος δεδομένων όπου η συνάρτηση παρουσιάζει περιοδικότητα. Σε αυτή την περίπτωση χρησιμοποιούνται και οι rational συναρτήσεις.

Τα δεδομένα που χρησιμοποιήθηκαν για τα πειράματα της παρούσας διπλωματικής εργασίας δεν παρουσιάζον περιοδικότητα. Για τις μετρήσεις που έχουν συλλεχθεί, είναι δύσκολο να προκύψει η παράγωγος της συνάρτησης. Γι'αυτούς τους λόγους θα μελετήσουμε μόνο τις μεθοδολογίες που μπορούν να εφαρμοστούν στα δεδομένα μας, δηλαδή τα πολυώνυμα Lagrange και τη Cubic Spline.

<u>Πολυώνυμα Lagrange</u>

Η πολυωνυμική μέθοδος Lagrange είναι μια γενίκευση της γραμμικής μεθόδου που περιγράφθηκε συνοπτικά στην προηγούμενη ενότητα. Η γραμμική συνάρτηση που υπολογίσαμε θεωρείται μια πολυωνυμική συνάρτηση πρώτου βαθμού. Για να γενικεύσουμε την έννοια της γραμμικής παρεμβολής / προεκβολής, θεωρούμε ότι έχουμε να κατασκευάσουμε ένα πολυώνυμο n βαθμού το οποίο πρέπει να περνά από τα n + 1 σημεία:

$(x_0, f(x_0)) , (x_1, f(x_1)) , \dots , (x_n, f(x_n))$.



Εικόνα 9: Κατασκευή Συνάρτησης Πολυωνύμου n Βαθμού

Σε αυτή την περίπτωση θα κατασκευάσουμε για κάθε k = 0, 1, … , n, μια συνάρτηση $L_{n,k}$ (x) όπου θα έχει την ιδιότητα $L_{n,k}$ (x_i) = 0 όταν i ≠ k και $L_{n,k}$ (x_k) = 1. Για να ικανοποιήσουμε την ιδιότητα $L_{n,k}$ (x_i) = 0 για κάθε i ≠ k απαιτείται ο αριθμητής του $L_{n,k}$ (x) να ισούται με:

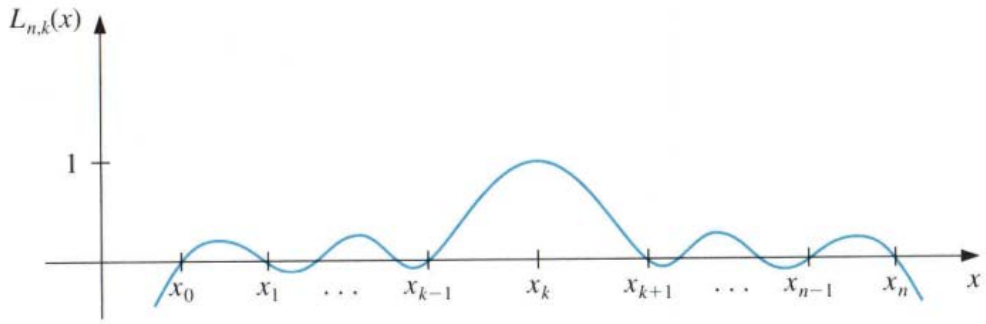
$$(x - x_0)\,(x - x_1) \cdot \cdot \cdot (x - x_{k-1})\,(x - x_{k+1}) \cdot \cdot \cdot (x - x_n)$$

Για να ικανοποιήσουμε την ιδιότητα $L_{n,k}$ (x_k) = 1 πρέπει ο παρονομαστής του $L_{n,k}$ (x) να ισούται με τον όρο που έχει εκτιμηθεί στο σημείο $x = x_k$.

Άρα:

$$L_{n,k}(x) = \frac{(x - x_0)(x - x_1)\ldots(x - x_{k-1})(x - x_{k+1})\ldots(x - x_n)}{(x_k - x_0)(x_k - x_1)\ldots(x_k - x_{k-1})(x_k - x_{k+1})\ldots(x_k - x_n)}$$

Η αναπαράσταση μιας τυπικής γραφικής παράστασης $L_{n,k}$ φαίνεται στο σχήμα που ακολουθεί.



Εικόνα 10: Γραφική Παράσταση Πολυωνύμου Lagrange

Οι μεθοδολογίες παρεμβολής και προεκβολής μπορούν εύκολα να εφαρμοστούν όταν υπολογιστεί το πολυώνυμο $L_{n,k}$. Συνοψίζοντας, η πολυωνυμική συνάρτηση δίνεται από τον τύπο:

$$P(x) = f(x_0)\, L_{n,0}(x) + \ldots + f(x_n)\, L_{n,n}(x) = \sum_{k=0}^{n} f(x_k)\, L_{n,k}(x)$$

όπου για κάθε k = 0,1, … , n,

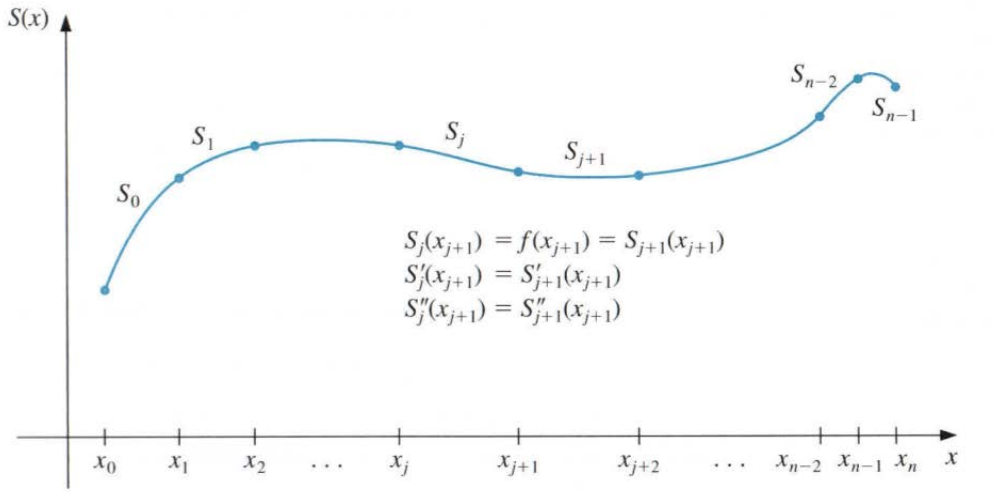
$$L_{n,k}(x) = \frac{(x - x_0)(x - x_1)\ldots(x - x_{k-1})(x - x_{k+1})\ldots(x - x_n)}{(x_k - x_0)(x_k - x_1)\ldots(x_k - x_{k-1})(x_k - x_{k+1})\ldots(x_k - x_n)}$$

$$= \prod_{\substack{i=o \\ i \neq k}}^{n} \frac{(X - X_i)}{(X_k - X_i)}$$

[21]

Cubic Splines

Για μεγαλύτερη ακρίβεια στα αποτελέσματα των πολυωνυμικών προσεγγίσεων χρησιμοποιούνται κυρίως κυβικά πολυώνυμα ανάμεσα στα ζεύγη των σημείων (δηλαδή ο βαθμός του πολυωνύμου ισούται με 3). Η μεθοδολογία αυτή ονομάζεται Cubic Spline Interpolation / Extrapolation [21]. Ένα τυπικό κυβικό πολυώνυμο αποτελείται από τέσσερις σταθερές ώστε να παρέχει την απαραίτητη ευελιξία στην διαδικασία σχηματισμού της καμπύλης. Ωστόσο, κατά την διαδικασία κατασκευής του cubic spline πολυωνύμου δεν είναι υποχρεωτικό η καμπύλη να περνάει από όλα τα σημεία που έχουν μετρηθεί.



Εικόνα 11: Γραφική Παράσταση Cubic Spline

Δεδομένης μια συνάρτησης f που ορίζεται στο διάστημα [a,b] και ενός συνόλου από σημεία $a = x_0 < x_1 < \ldots < x_n = b$, η cubic spline παρεμβολή / προεκβολή S για την f είναι μια συνάρτηση που ικανοποιεί τις παρακάτω συνθήκες.

a. S(x) είναι ένα πολυώνυμο τρίτου βαθμού, που ορίζεται ως $S_j(x)$ στο υποδιάστημα $[x_j , x_{j+1}]$ για κάθε j = 0, 1, …, n-1.
b. $S_j(x_j) = f(x_j)$ και $S_j(x_{j+1}) = f(x_{j+1})$ για κάθε j = 0, 1, …, n-1.
c. $S_{j+1}(x_{j+1}) = S_j(x_{j+1})$ για κάθε j = 0, 1, …, n-2.
d. $S'_{j+1}(x_{j+1}) = S'_j(x_{j+1})$ για κάθε j = 0, 1, …, n-2.
e. $S''_{j+1}(x_{j+1}) = S''_j(x_{j+1})$ για κάθε j = 0, 1, …, n-2.
f. Ικανοποιείται ένας από τους παρακάτω περιορισμούς:
   i. $S''(x_0) = S''(x_n) = 0$ (free / natural boundary)
   ii. $S'(x_0) = f'(x_0)$ και $S'(x_n) = f'(x_n)$ (clumped boundary)

Όταν συναντάμε την πρώτη περίπτωση περιορισμού (free boundary) η καμπύλη ονομάζεται natural spline και η γραφική της παράσταση έχει το σχήμα μιας μακριάς, ελαστικής ράβδου που αναγκάζεται να περάσει από όλα τα σημεία $\{ (x_0,f(x_0)), (x_1,f(x_1)) , \ldots , (x_n,f(x_n)) \}$. Ο δεύτερος περιορισμός (clumped boundary) οδηγεί σε πιο ακριβείς προσεγγίσεις αφού εμπεριέχει περισσότερη πληροφορία της συνάρτησης. Ωστόσο, για να ισχύσει αυτός ο περιορισμός, είναι απαραίτητο να υπάρχουν τιμές στα παράγωγα των τελικών σημείων ή να υπάρχει ακριβής προσέγγιση των τιμών τους.

Για την κατασκευή της συνάρτησης cubic spline παρεμβολής / προεκβολής δεομένης μια συνάρτησης f, εφαρμόζουμε τις παραπάνω συνθήκες στο πολυώνυμο τρίτου βαθμού:

$$S_j(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3 \text{ , για κάθε } j = 0, 1, \ldots, n - 1$$

Από την συνθήκη (c) ισχύει ότι $S_j(x_j) = a_j = f(x_j)$, οπότε αντικαθιστούμε:

$$a_{j+1} = S_{j+1}(x_{j+1}) = S_j(x_{j+1}) = a_j + b_j(x_{j+1} - x_j) + c_j(x_{j+1} - x_j)^2 + d_j(x_{j+1} - x_j)^3 ,$$

για κάθε j = 0, 1, …, n - 2.

Επειδή ο όρος $(x_{j+1} - x_j)$ επαναλαμβάνεται συνεχώς, θα ορίσουμε μια νέα μεταβλητή h_i για συντομογραφία. Δηλαδή,

$$h_i = (x_{j+1} - x_j), \text{ για κάθε } j = 0, 1, \ldots, n-1$$

Αν ορίσουμε $a_n = f(x_n)$, τότε η εξίσωση γίνεται:

$$a_{j+1} = a_j + b_j h_j + c_j h_j^2 + d_j h_j^3 , \qquad \text{για κάθε } j = 0, 1, \ldots, n-1 \qquad (1)$$

Με τον ίδιο τρόπο ορίζουμε $b_n = S'(x_n)$, και παρατηρούμε ότι:

$$S_j'(x) = b_j + 2c_j(x - x_j) + 3d_j(x - x_j)^2$$

συνεπάγεται $S_j'(x_j) = b_j$, για κάθε j = 0, 1, …, n - 1

Εφαρμόζοντας την συνθήκη (d) έχουμε:

$$b_{j+1} = b_j + 2c_j h_j + 3d_j h_j^2 , \text{ για κάθε } j = 0, 1, \ldots, n-1 \qquad (2)$$

Μια άλλη σχέση μεταξύ των συντελεστών της S_j εξασφαλίζεται με τον ορισμό του $c_n = S''(x_n) / 2$ και την εφαρμογή της συνθήκης (c). Έτσι για κάθε j = 0, 1, …, n – 1,

$$c_{j+1} = c_j + 3d_j h_j \qquad (3)$$

Επιλύοντας ως προς d_j στην (3) και αντικαθιστώντας αυτή την τιμή στις (1) & (2) έχουμε για κάθε j = 0, 1, …, n – 1, τις παρακάτω νέες εξισώσεις:

$$a_{j+1} = a_j + b_j h_j + (h_j^2 / 3) \cdot (2c_j + c_{j+1}) \qquad (4)$$

και

$$b_{j+1} = b_j + h_j \cdot (c_j + c_{j+1}) \qquad (5)$$

Αν επιλύσουμε την (4) πρώτα ως προς b_j έχουμε

$$b_j = (1 / h_j) \cdot (a_{j+1} - a_j) - (h_j / 3) \cdot (2c_j + c_{j+1}) \qquad (6)$$

και έπειτα με αναγωγή ως προς b_{j-1} ,

$$b_{j-1} = (1 / h_{j-1}) \cdot (a_j - a_{j-1}) - (h_{j-1} / 3) \cdot (2c_{j-1} + c_j)$$

Αντικαθιστώντας αυτές τις τιμές στην εξίσωση (5) προκύπτει η γραμμική εξίσωση:

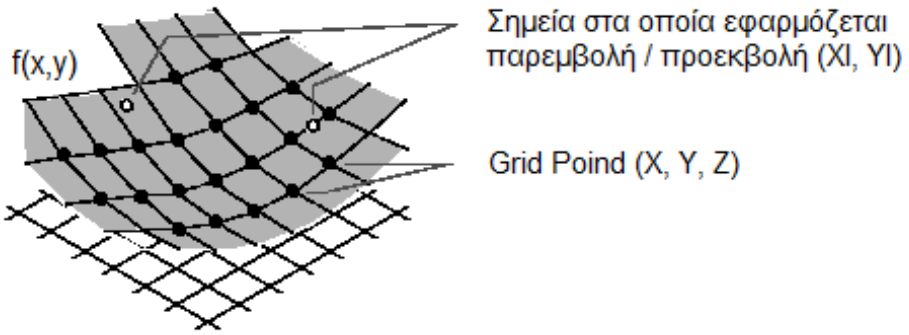
$$h_{j-1} c_{j-1} + 2(h_{j-1} + h_j) c_j + h_j c_{j+1} = (3 / h_j) \cdot (a_{j+1} - a_j) - (3 / h_{j-1}) \cdot (a_j - a_{j-1})$$

για κάθε j = 0, 1, …, n – 1 (7)

Σε αυτή την εξίσωση ο μόνος άγνωστος είναι ο $\{ c_j \}^{n}_{j=0}$ αφού οι τιμές των $\{ h_j \}^{n-1}_{j=0}$ και $\{ a_j \}^{n}_{j=0}$ είναι ήδη γνωστές από τις αρχικές μετρήσεις $\{ x_j \}^{n}_{j}$ και τις αντίστοιχες τιμές τους. Μόλις καθοριστούν οι τιμές των $\{ c_j \}^{n}_{j=0}$ είναι πολύ εύκολο να βρεθούν οι τιμές των $\{ b_j \}^{n1}_{j=0}$ από την (6) και των $\{ d_j \}^{n-1}_{j=0}$ από την (3) και έτσι να κατασκευαστεί το κυβικό πολυώνυμο $\{ S_j (x) \}^{n-1}_{j=0}$.

3.3 Εργαλεία (matlab toolbox)

Για την εφαρμογή της μεθοδολογίας extrapolation στα πειράματα της διπλωματικής εργασίας, χρησιμοποιήθηκε το εργαλείο Matlab 7.11.0 (R2010b). Το Matlab παρέχει τη συνάρτηση interp2 που εφαρμόζεται σε δεδομένα δύο διαστάσεων και υλοποιεί τις μεθόδους παρεμβολής και προεκβολής [34]. Η εντολή interp2 παρεμβάλλει ανάμεσα στα σημεία δεδομένων (Εικόνα 12). Βρίσκει τις τιμές της δισδιάστατης συνάρτησης f(x,y) βασιζόμενη στις τιμές των σημείων του πλέγματος (grid points).



Εικόνα 12: Η εντολή interp2 (Matlab)

Η σύνταξη της εντολής είναι:

Z_i = interp2 (X, Y, Z, XI, YI, method)

Τα διανύσματα X και Y περιλαμβάνουν τις μετρήσεις για τις δύο διαστάσεις των δεδομένων. Οι X και Y καθορίζουν τα σημεία για τα οποία δίνονται τα δεδομένα στον πίνακα Z. Συνεπώς, αν X = 1:n και Y = 1:m, τότε το μέγεθος του πίνακα Z θα πρέπει να είναι [m,n].

Τα XI και YI καθορίζουν το σημείο που θέλουμε να προβλέψουμε την τιμή του. Τα XI, YI μπορεί να είναι και διανύσματα. Σε αυτή την περίπτωση η εντολή interp2 επιστρέφει τις τιμές του Z για τα σημεία (XI(i,j),YI(i,j)). Εναλλακτικά, μπορούμε να ορίσουμε αριθμητικά την γραμμή και την στήλη για τους άξονες X και Y αντίστοιχα.

Το όρισμα method καθορίζει τι είδους παρεμβολή θα χρησιμοποιηθεί. Οι τιμές που μπορεί να πάρει είναι οι εξής:

- 'nearest': Παρεμβολή πλησιέστερου γείτονα (Nearest neighbour interpolation)
- 'linear' : Γραμμική Παρεμβολή (Linear interpolation)

- 'spline': Cubic spline Παρεμβολή
- 'cubic': Cubic παρεμβολή. Εφαρμόζεται μόνο σε δεδομένα που είναι ομοιόμορφα κατανεμημένα. Διαφορετικά είναι ίδια με την μέθοδο 'spline'.

Αν δεν ορίσουμε τι μέθοδο θα χρησιμοποιήσουμε τότε εξορισμού επιλέγεται η γραμμική παρεμβολή.

Για να εφαρμόσουμε προεκβολή στα δεδομένα μας, πρέπει να ορίσουμε ως μέθοδο τη 'spline'. Διαφορετικά το αποτέλεσμα που προκύπτει για τις τιμές εκτός των ορίων είναι Nan (not a number).

4 ΕΚΤΙΜΗΣΗ ΜΕΜΟΝΩΜΕΝΩΝ ΤΙΜΩΝ

4.1 Περιγραφή προβλήματος και εφαρμογές

Τα περισσότερα σύνολα δεδομένων εμπεριέχουν ελλιπείς τιμές. Οι ελλιπείς τιμές υποδεικνύονται από εγγραφές που η τιμή τους βρίσκεται εκτός των προκαθορισμένων ορίων, ή από αρνητικούς αριθμούς σε πεδία που ορίζουν μόνο θετικές τιμές ή από την τιμή 0 σε αριθμητικά πεδία που δεν προβλέπεται μηδενική τιμή. Οι περισσότερες μέθοδοι μηχανικής μάθησης κάνουν την έμμεση παραδοχή ότι το γεγονός της έλλειψης μιας τιμής ενός γνωρίσματος σε μια μέτρηση δεν έχει κάποια ιδιαίτερη σημασία (δηλαδή απλώς δεν είναι γνωστή η τιμή). Ωστόσο μπορεί να υπάρχει κάποιος σημαντικός λόγος που η τιμή λείπει και αυτό να μας οδηγήσει σε κάποια πληροφορία σχετικά με το ίδιο το γνώρισμα. Αν ισχύει αυτό, είναι προτιμότερο η τιμή του γνωρίσματος να καταγράφεται ως “μη δοκιμασμένη” (not tested) ή ως ένα διαφορετικό γνώρισμα στο σύνολο των δεδομένων. Όπως έχουν δείξει αρκετές περιπτώσεις παραδειγμάτων, μόνο κάποιος που είναι πολύ εξοικειωμένος με τα δεδομένα μπορεί να διαμορφώσει τεκμηριωμένη γνώμη για το αν μια συγκεκριμένη τιμή που λείπει έχει κάποια επιπλέον σημασία ή αν πρέπει απλά να κωδικοποιηθεί ως μια συνηθισμένη ελλιπής τιμή. Φυσικά, εάν υπάρχουν διάφοροι τύποι ελλιπών τιμών, είναι εκ πρώτης όψεως αποδεικτικό στοιχείο ότι συμβαίνει κάτι που πρέπει να διερευνηθεί περεταίρω [26].

Ένα σύνολο δεδομένων που συλλέγεται από αισθητήρες είναι πολύ μεγάλο σε όγκο λόγω της συνεχούς ενημέρωσης του. Σε αυτή την περίπτωση είναι αδύνατον να μελετηθούν οι λόγοι που λείπει μια μεμονωμένη τιμή. Γι’ αυτό το λόγο πρέπει να δημιουργηθεί ένα μοντέλο πρόβλεψης τιμών που να αντιπροσωπεύει τη πλειονότητα των δεδομένων και να μπορεί να ικανοποιήσει σε γενικό επίπεδο όλες τις μεμονωμένες ελλιπής τιμές. Σε αυτό το κεφάλαιο θα μελετηθούν αλγόριθμοι κατηγοριοποίησης (classification algorithms) που κατασκευάζουν δένδρα απόφασης για να χρησιμοποιηθούν ως μοντέλα πρόβλεψης ελλιπών τιμών.

4.2 Αλγόριθμοι Κατηγοριοποίησης

Στην μηχανική μάθηση, το πρόβλημα της κατηγοριοποίησης (classification) είναι πολύ σύνηθες. Κατηγοριοποίηση των δεδομένων είναι η διαδικασία κατά την οποία εξετάζονται τα βασικότερα χαρακτηριστικά ενός συνόλου δεδομένων και βάσει αυτών κατηγοριοποιείται ένα νέο αντικείμενο σε μια προκαθορισμένη κατηγορία σύμφωνα με το μοντέλο κατηγοριοποίησης. Για την δημιουργία του μοντέλο κατηγοριοποίησης πρέπει πρώτα να δοθεί ένα σύνολο με δεδομένα εκμάθησης (training data) για την κάθε κατηγορία. Έπειτα, με την χρήση αυτού του μοντέλου θα μπορούν να κατηγοριοποιηθούν δεδομένα τα οποία δεν ανήκουν σε κάποια κατηγορία (άγνωστα / νέα αντικείμενα). Η κατηγοριοποίηση των νέων δεδομένων μπορεί να δείξει μια μεταβλητότητα στην συμπεριφορά του προβλήματος που εξετάζεται και βάσει αυτών να εξαχθούν στατιστικά αποτελέσματα.

Τα δένδρα απόφασης (decision trees) χρησιμοποιούνται αρκετά στους κλάδους της στατιστικής, της εξόρυξης γνώσης και της μηχανικής μάθησης, ως μεθοδολογίες κατηγοριοποίησης. Στόχος είναι να δημιουργηθεί ένα μοντέλο που θα μπορεί να προβλέπει την τιμή μιας μεταβλητής βασιζόμενο στην είσοδο δεδομένων των υπολοίπων μεταβλητών.

Τα δέντρα που χρησιμοποιούνται για την πρόβλεψη τιμών είναι ακριβώς όπως τα συνηθισμένα δένδρα απόφασης εκτός από το γεγονός ότι σε κάθε φύλλο αποθηκεύουν είτε μια τιμή (class value) που αντιπροσωπεύει τη μέση τιμή των αντικειμένων (μετρήσεων) που φτάνουν στο συγκεκριμένο φύλλο, ή ένα γραμμικό μοντέλο παλινδρόμησης που προβλέπει την τιμή των αντικειμένων που φτάνουν το φύλλο. Στην πρώτη περίπτωση το δένδρο ονομάζεται δένδρο παλινδρόμησης (regression tree) ενώ στην δεύτερη περίπτωση πρότυπο δένδρο (model tree) [26].

Τα regression trees και model trees δημιουργούνται χρησιμοποιώντας πρώτα έναν αλγόριθμο δένδρου απόφασης για να κατασκευαστεί το αρχικό δένδρο. Ωστόσο, ενώ οι περισσότεροι αλγόριθμοι κατασκευής δένδρων απόφασης επιλέγουν το χαρακτηριστικό διάσπασης (splitting attribute) με βάση την μεγιστοποίηση του κέρδους πληροφορίας (information gain), τα δένδρα που χρησιμοποιούνται για την πρόβλεψη τιμών στοχεύουν στην ελαχιστοποίηση των διακυμάνσεων κάθε εσωτερικού υποσυνόλου που βρίσκεται κάτω από κάθε κόμβο. Μόλις διαμορφωθεί το βασικό δέντρο, δίνεται έμφαση στο “κλάδεμα” (pruning) του δέντρου ξεκινώντας από κάθε φύλλο και πηγαίνοντας προς τα πάνω, όπως συμβαίνει και με τα απλά δέντρα απόφασης. Η μόνη διαφορά μεταξύ του regression tree και του model tree είναι ότι στο δεύτερο, κάθε κόμβος αντικαθίσταται από μια παλινδρομική συνάρτηση αντί για μια σταθερά. Τα αντικείμενα που χρησιμοποιούνται για να ορισθεί αυτή η παλινδρομική συνάρτηση είναι ακριβώς τα ίδια που συμμετέχουν στη λήψη αποφάσεων για τα υπο-δένδρα που θα “κλαδευτούν”, δηλαδή τους κόμβους κάτω από τον τρέχον.

Σε αυτή την ενότητα θα μελετηθούν οι κυριότεροι αλγόριθμοι κατασκευής δένδρων απόφασης που αφορούν τόσο regression trees όσο και model trees.

4.2.1 ID3 / C4.5

Ο ID3 είναι ένας αλγόριθμος μάθησης που υλοποιείται με ένα απλό δένδρο απόφασης και αναπτύχτηκε από τον Ross Quinlan (1983) [27]. Ο ID3 είναι ο πρόδρομος του C4.5. Η βασική ιδέα του αλγορίθμου είναι η κατασκευή ενός δένδρου από πάνω προς τα κάτω, εφαρμόζοντας «άπληστη» αναζήτηση στο σύνολο δεδομένων που δίνεται. Δηλαδή για να τοποθετηθεί ένα αντικείμενο σε έναν κόμβο γίνεται έλεγχος όλων των κόμβων του δένδρου. Για να επιλεχθεί το χαρακτηριστικό γνώρισμα που θα χρησιμοποιηθεί για την κατηγοριοποίηση (classification) του συνόλου των δεδομένων, εισάγεται μία μετρική που ονομάζεται κέρδος πληροφορίας (information gain).

Για να βρεθεί ο βέλτιστος τρόπος κατηγοριοποίησης των δεδομένων πρέπει να μειωθεί το βάθος του δένδρου. Δηλαδή να ελαχιστοποιηθούν οι ερωτήσεις που γίνονται σε κάθε κόμβο για την επιλογή του επόμενου μονοπατιού. Η συνάρτηση αυτή, που μπορεί να μετρήσει ποιες ερωτήσεις παρέχουν την πιο ισορροπημένη διάσταση, ονομάζεται information gain. Για να οριστεί με ακρίβεια το information gain πρέπει πρώτα να οριστεί η έννοια της εντροπίας.

Έστω ότι έχουμε ένα σύνολο S για κατηγοριοποίηση, όπου περιέχει θετικά και αρνητικά παραδείγματα. Η εντροπία του S δίνεται από τον τύπο:

$$E(S) = - P(positive) \log_2 P(positive) - P(negative) \log_2 P(negative)$$

όπου:

P(positive): το ποσοστό των θετικών παραδειγμάτων στο S

P(negative): το ποσοστό των αρνητικών παραδειγμάτων στο S

Άρα ο γενικότερος τύπος της εντροπίας ορίζεται ως:

$$E(S) = -\sum_{j=1}^{n} f_S(j) \log_2 f_S(j)$$

όπου:

n: ο αριθμός των διαφορετικών τιμών των γνωρισμάτων στο σύνολο δεδομένων S (η εντροπία υπολογίζεται για ένα επιλεγμένο γνώρισμα)

$f_S(j)$: το ποσοστό της j τιμής στο σύνολο δεδομένων S

Ο αλγόριθμος του ID3 μπορεί να συνοψιστεί στα εξής τρία βήματα:

1. Υπολόγισε την εντροπία για κάθε ένα γνώρισμα του συνόλου των δεδομένων.
2. Επέλεξε το γνώρισμα που ελαττώνει την εντροπία (ή ισοδύναμα, αυξάνει το information gain).
3. Δημιούργησε έναν νέο κόμβο με αυτό το γνώρισμα.

Ένας κόμβος με εντροπία ίση με το 0 αποτελεί φύλλο του δένδρου. Διαφορετικά ο κόμβος αυτός χρειάζεται να ξαναδιασπαστεί για να μπορέσει να ταξινομήσει τα δεδομένα. Ο αλγόριθμος ID3 εκτελείται αναδρομικά σε όλους τους κόμβους που δεν είναι φύλλα μέχρι να ταξινομηθούν όλα τα δεδομένα.

Ο αλγόριθμος C4.5 αναπτύχθηκε και αυτός από τον Ross Quinlan και ουσιαστικά αποτελεί μια επέκταση του αλγορίθμου ID3. Κατασκευάζει δένδρα απόφασης από ένα εκπαιδευτικό σύνολο δεδομένων με τον ίδιο τρόπο που κατασκευάζει και ο ID3, χρησιμοποιώντας την έννοια του information gain. Ωστόσο, ο C4.5 έχει αρκετές βελτιστοποιήσεις σε σχέση με τον ID3. Αυτές οι βελτιώσεις περιλαμβάνουν μεθόδους για την αντιμετώπιση των αριθμητικών γνωρισμάτων, των ελλιπών τιμών, του «θορύβου» στα δεδομένα και τη δημιουργία κανόνων από το δένδρο που δημιουργείται [26].

Πιο συγκεκριμένα, ο C4.5 μπορεί να χειριστεί δεδομένα και χαρακτηριστικά τόσο με διακριτά όσο και με συνεχή πεδία τιμών. Στην περίπτωση που υπάρχουν συνεχή πεδία τιμών (δηλαδή αριθμητικά δεδομένα), ο αλγόριθμος δημιουργεί ένα όριο (threshold) και έπειτα χωρίζει τα δεδομένα σε εκείνα που έχουν τιμή πάνω από το όριο και εκείνα που η τιμή τους είναι μικρότερη ή ίση του ορίου. Για τον χειρισμό των δεδομένων εκμάθησης με ελλιπείς τιμές, επιτρέπει την χρήση του συμβόλου " ? " όπου λείπει η τιμή. Οι τιμές που λείπουν δεν χρησιμοποιούνται στους υπολογισμούς τις εντροπίας και του information gain. Όσον αφορά την μείωση του θορύβου στα δεδομένα, ο αλγόριθμος C4.5 εφαρμόζει μια τεχνική εκτίμησης ποσοστού λάθους πάνω στα ίδια τα δεδομένα εκπαίδευσης. Η βασική ιδέα είναι να εξεταστούν όλα τα αντικείμενα που φθάνουν σε κάθε κόμβο. Κάθε κόμβος αντιπροσωπεύεται από την πλειοψηφία των αντικειμένων. Άρα, από το σύνολο των αντικειμένων κάθε κόμβου ένα ποσοστό αποτελεί σφάλμα. Αυτό το σφάλμα υπολογίζεται με την χρήση της

μεθοδολογίας Bernoulli. Τέλος, από το δένδρο που κατασκευάζεται, είναι εύκολο να δημιουργηθούν κανόνες που να το περιγράφουν. Για κάθε κόμβο έχει δημιουργηθεί ένας κανόνας. Με την σύζευξη όλων των κανόνων ενός μονοπατιού από τη ρίζα έως το φύλλο δημιουργούμε τους κανόνες του δένδρου. Οι κανόνες που παράγονται, δεν αφήνουν περιθώρια παρερμηνείας ανεξάρτητα από την σειρά που θα εκτελεστούν.

4.2.2 M5P

Ο αλγόριθμος M5P είναι μια παραλλαγή του αλγορίθμου M5 που παρουσιάστηκε αρχικά από τον J. R. Quinlan [23] και δημιουργεί δένδρα αποφάσεων από πρότυπα μοντέλα (model trees). Ο M5P συνδυάζει ένα συμβατικό δέντρο απόφασης με τη δυνατότητα ύπαρξης γραμμικής συνάρτησης στους κόμβους του. Ο M5 αλγόριθμος είναι ένας από τους πιο συχνά χρησιμοποιούμενους ταξινομητές (classifiers) του είδους του. Τα φύλλα των δένδρων που χτίζει (model trees), αποτελούνται από γραμμικά μοντέλα πολλών μεταβλητών και οι κόμβοι του δένδρου επιλέγονται βάσει του γνωρίσματος που θα ελαχιστοποιήσει το αναμενόμενο σφάλμα σε συνάρτηση με την τυπική απόκλιση της τιμής εξόδου. [24]

Για την εφαρμογή του αλγορίθμου δημιουργείται αρχικά ένα δένδρο απόφασης με βάση το training set, χρησιμοποιώντας έναν αλγόριθμο κατασκευής δένδρων αποφάσεων. Όμως, ως κριτήριο διαχωρισμού κάθε εσωτερικού κόμβου δεν χρησιμοποιείται η μεγιστοποίηση του information gain. Το κριτήριο είναι η ελαχιστοποίηση της διακύμανσης του υποσυνόλου της τάξης κάτω από κάθε εσωτερικό κόμβο. Η διαδικασία διαχωρισμού του M5P αλγορίθμου σταματά όταν για κάθε τιμή της τάξης, τα νέα δεδομένα που φτάνουν διαφέρουν κατά πολύ λίγο ή όταν δεν υπάρχουν άλλα δεδομένα. Έτσι, για κάθε νέο κόμβο του δένδρου που δημιουργείται, έχει υπολογιστεί ένα γραμμικό μοντέλο (με την χρήση της γραμμικής παλινδρόμησης).

Το επόμενο βήμα είναι η απλοποίηση του παλινδρομικού δένδρου που δημιουργήθηκε στο προηγούμενο βήμα. Το δένδρο «κλαδεύεται», διαγράφοντας τους κόμβους των γραμμικών μοντέλων, των οποίων τα γνωρίσματα δεν αυξάνουν το σφάλμα.

Τέλος, για να αποφευχθούν οι απότομες ασυνέχειες των δεδομένων μεταξύ των υπο-δένδρων, εφαρμόζεται μια διαδικασία εξομάλυνσης. Σύμφωνα με αύτη την διαδικασία, εξομαλύνεται κάθε ένας κόμβος συνδυάζοντας το μοντέλο πρόβλεψης που προκύπτει κατά μήκος της διαδρομής με την τιμή πρόβλεψης του γραμμικού μοντέλου του συγκεκριμένου κόμβου. Έτσι απαλείφονται οι κόμβοι που έχουν μεγαλύτερο σφάλμα ταξινόμησης (classification error) από αυτό που προκύπτει από το γραμμικό μοντέλο των ενδιάμεσων κόμβων. Σκοπός του τελευταίου βήματος είναι να μειωθεί το μέγεθος του δένδρου χωρίς να ελαττωθεί η ακρίβειά του.

Ο αλγόριθμος M5P παράγει μοντέλα από δεδομένα που περιλαμβάνουν συνεχείς τάξεις. Επίσης μπορεί να χειριστεί αποτελεσματικά απαριθμημένα γνωρίσματα αλλά και ελλιπείς τιμές των γνωρισμάτων. Όλα τα γνωρίσματα μετατρέπονται σε δυαδικές μεταβλητές έτσι ώστε όλοι οι διαχωρισμοί του αλγορίθμου να είναι δυαδικοί. Όταν δεν υπάρχει τιμή στο γνώρισμα και δεν μπορεί να γίνει ο διαχωρισμός, χρησιμοποιείται μια τεχνική που ονομάζεται “surrogate splitting”. Σε αυτή την περίπτωση επιλέγεται ένα άλλο γνώρισμα να αντικαταστήσει το αρχικό. Το γνώρισμα που επιλέγεται είναι αυτό που έχει την μεγαλύτερη συσχέτιση με το αρχικό. Όταν τελειώσει η διαδικασία χωρισμού, όλες οι τιμές που έλλειπαν αντικαθιστούνται από την μέση τιμή του training set του αντίστοιχου γνωρίσματος. Κατά την διάρκεια των δοκιμών

(tests), μια ελλιπής τιμή αντικαθίσταται από την μέση τιμή του γνωρίσματος, από όλα τα δεδομένα του training set που έχουν ταξινομηθεί σε αυτό τον κόμβο, με αποτέλεσμα την επιλογή του πιο δημοφιλούς υπο-κόμβου [25].

4.2.3 RepTree

Ο REPTree είναι ένας γρήγορος αλγόριθμος μάθησης που χρησιμοποιεί δένδρα απόφασης. Χτίζει ένα regression tree χρησιμοποιώντας τη μετρική του information gain και τη μείωση της διακύμανσης στους εσωτερικούς κόμβους. Εφαρμόζεται μόνο σε αριθμητικά δεδομένα και δίνει ως αποτέλεσμα την μέση τιμή των αντικειμένων του κάθε φύλλου (δηλαδή επιστρέφει πάλι αριθμητικό αποτέλεσμα). Οι ελλιπείς τιμές αντιμετωπίζονται με τη χρήση ενός ορίου (threshold), όπως ακριβώς γίνεται και στον αλγόριθμο C4.5 [26].

4.2.4 Decision Stump

Ο όρος Decision Stump αναφέρεται σε ένα μοντέλο μηχανικής μάθησης που αποτελείται από ένα δένδρο απόφασης ενός επιπέδου (βάθους 1). Δηλαδή, το δένδρο αποτελείται από έναν εσωτερικό κόμβο (την ρίζα του δένδρου) ο οποίος καταλήγει άμεσα στους τερματικούς κόμβους (τα φύλλα του δένδρου). Το decision stump δένδρο κάνει τις προβλέψεις βασιζόμενο στην τιμή μίας μόνο μεταβλητής εισόδου. Πολλές φορές το decision stump δένδρο αποκαλείται και ως δένδρο “ενός κανόνα” (1 - rule).

Ένας απλός αλγόριθμος “ενός επιπέδου” επιλέγει ένα γνώρισμα για να προβλέψει την κατηγορία της τιμής που λείπει. Έστω ότι έχουμε ένα ενιαίο συσχετισμένο χαρακτηριστικό, A^0 και ένα σύνολο από q ασυσχέτιστα γνωρίσματα A_1, A_2, …, A_q. Το γνώρισμα A_i θεωρείται ασυσχέτιστο όταν ο εννοιολογικός στόχος (δηλαδή το A^0) δεν περιέχει το A_i. Για λόγους απλοποίησης, θα θεωρήσουμε ότι οι μετρήσεις απεικονίζονται ως ένα σύνολο Boolean μεταβλητών.

Ο αλγόριθμος “ενός επιπέδου” υπολογίζει ένα σκορ, για κάθε ένα γνώρισμα A, καταμετρώντας πόσο καλά χωρίζει το γνώρισμα A το σύνολο της τάξης. Αφού θεωρήσαμε ότι οι έννοιες και τα γνωρίσματα είναι Boolean, μπορούμε απλώς να μετρήσουμε πόσες φορές η τιμή ανάμεσα στην έννοια και το γνώρισμα είναι η ίδια (είτε αληθής είτε ψευδής) και πόσες φορές διαφέρει, σε ένα training set μεγέθους n. Όταν είναι ίδια αναφερόμαστε ως |A=C| ενώ όταν διαφέρουν ως |A≠C|. Άρα το σκορ υπολογίζεται από τον παρακάτω τύπο:

$$score(A) = \max(|A=C|, |A\neq C|) / n$$

και έχει εύρος τιμών: $½ \leq score(A) \leq 1$.

Ο αλγόριθμος επιλέγει το γνώρισμα με το καλύτερο σκορ. Σε περίπτωση ισοβαθμίας επιλέγεται ένα από τα γνωρίσματα με το καλύτερο σκορ.

Σκοπός είναι να αυξήσουμε την πιθανότητα της σωστής κατηγοριοποίησης των νέων δεδομένων από το decision stump, έπειτα από n εκπαιδευτικά δεδομένα [22]. Οι παράγοντες που επηρεάζουν αυτή την πιθανότητα είναι: (i) το πλήθος των

ασυσχέτιστων γνωρισμάτων, (ii) το πλήθος των τάξεων και ο θόρυβος των γνωρισμάτων, (iii) οι συχνότητες των τάξεων και των γνωρισμάτων και (iv) το μέγεθος του εκπαιδευτικού δείγματος (training set)

4.2.5 Πολυπλοκότητα Επαγωγής Δένδρου

Αφού μελετήθηκαν τα βασικά σημεία της δημιουργίας των δένδρων απόφασης, θα εξετάσουμε την πολυπλοκότητα της επαγωγής των δένδρων [26]. Έστω ότι έχουμε n δεδομένα εκπαίδευσης και m διαφορετικά χαρακτηριστικά βάσει των οποίων θα γίνει η κατηγοριοποίηση. Το βάθος ενός δένδρου είναι της τάξης log n, δηλαδή O(log n). Αυτός είναι ο κανονικός ρυθμός ανάπτυξης ενός δέδνρου. Το υπολογιστικό κόστος κατασκευής τους δένδρου είναι O(m·n·log n).

Τώρα, αν ένα δένδρο έχει αριθμητικά χαρακτηριστικά, αυτά πρέπει να ταξινομηθούν. Όταν γίνει η αρχική ταξινόμηση δεν χρειάζεται επιπλέον ταξινόμηση στα υπόλοιπα βήματα κάθε αλγορίθμου. Η πολυπλοκότητα της ταξινόμησης είναι O(n log n).

Έπειτα η διαδικασία του «κλαδέματος» και αντικατάστασης κάθε υποδένδρου είναι γραμμική. Στην αρχή γίνεται μια εκτίμηση του σφάλματος για κάθε κόμβο και ύστερα εξετάζουμε ποιος κόμβος χρειάζεται αντικατάσταση. Συνεπώς, επειδή κάθε δένδρο έχει το πολύ n φύλλα, η πολυπλοκότητα αντικατάστασης αυτών είναι O(n).

Τέλος, η διαδικασία ανύψωσης κάθε υποδένδρου έχει πολυπλοκότητα ίση με την αντικατάσταση του υποδένδρου. Ωστόσο, υπάρχει και ένα επιπλέον κόστος για την αναταξινόμηση των μετρήσεων. Κατά τη διάρκεια της διαδικασίας, κάθε μέτρηση μπορεί να αναταξινομηθεί στους κόμβους ανάμεσα στη ρίζα και τα φύλλα μέχρι και O(log n) φορές. Έτσι το συνολικό πλήθος ανακατατάξεων φτάνει στο O(n log n). Η αναταξινόμηση που γίνεται κοντά στη ρίζα του δένδρου είναι O(log n),ενώ όταν γίνεται σε ένα μέσο βάθος είναι περίπου η μισή. Άρα η συνολική πολυπλοκότητα της ανύψωσης του υποδένδρου είναι $O(n (\log n)^2)$.

Λαμβάνοντας υπόψιν όλες τις παραπάνω διαδικασίες, η συνολική πολυπλοκότητα επαγωγής του δένδρου είναι:

$O(m \cdot n \cdot \log n) + O(n (\log n)^2)$

4.3 Εργαλεία

4.3.1 WEKA

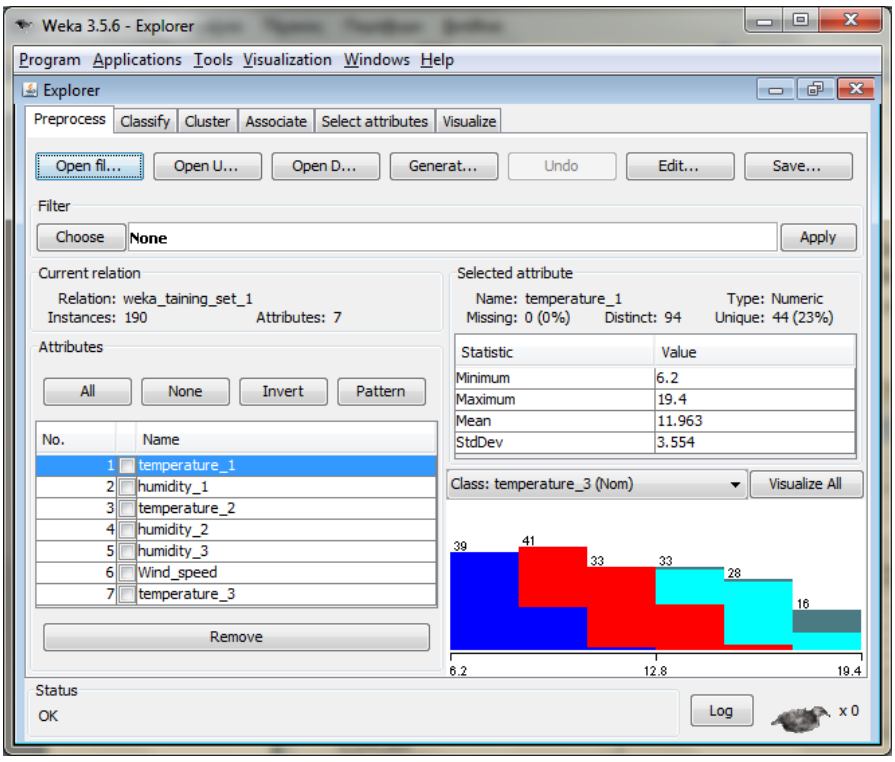
Το Weka (ακρωνύμιο του Waikato Environment for Knowledge Analysis) είναι ένα εργαλείο ελεύθερου λογισμικού, που αναπτύχθηκε από το Πανεπιστήμιο Waikato της Νέας Ζηλανδίας. Το Weka περιλαμβάνει μια μεγάλη συλλογή από υλοποιημένους αλγόριθμους μηχανικής μάθησης και εργαλεία προεπεξεργασίας δεδομένων. Παρέχει εκτενή υποστήριξη σε όλα τα στάδια των πειραμάτων για την εξόρυξη γνώσης, συμπεριλαμβανομένης της προετοιμασίας των δεδομένων εισόδου, την αξιολόγηση των στατιστικών μοντέλων που προκύπτουν και την απεικόνιση των δεδομένων εισόδου και των αποτελεσμάτων. Επίσης, το Weka περιλαμβάνει μεθοδολογίες για όλα τα γνωστά προβλήματα εξόρυξης δεδομένων: παλινδρόμηση (regression), κατηγοριοποίηση (classification), ομαδοποίηση (clustering), εξαγωγή κανόνων συσχέτισης (association rule mining) και επιλογή γνωρισμάτων (attribute selection).

Όλοι οι αλγόριθμοι που περιλαμβάνονται στο Weka παίρνουν ως είσοδο αρχεία ARFF (Εικόνα 13). Στην αρχή κάθε αρχείου ARFF (Attribute Relation File Format) ορίζεται η σχέση των δεδομένων (@relation όνομα_σχέσης) και αμέσως μετά πρέπει να οριστούν τα γνωρίσματα των δεδομένων (@attribute όνομα_γνωρίσματος). Για τα γνωρίσματα που παίρνουν λεκτικές τιμές πρέπει να οριστούν ποιες είναι οι επιτρεπτές τιμές μέσα σε άγκιστρο. Τα αριθμητικά γνωρίσματα πρέπει να οριστούν ως numeric. Αφού οριστούν η σχέση και τα γνωρίσματα ακολουθούν τα δεδομένα (@data), τα οποία πρέπει να ακολουθούν την ίδια σειρά με τα γνωρίσματα που ορίστηκαν προηγουμένως και να χωρίζονται μεταξύ τους με κόμμα ή με τον ειδικό χαρακτήρα tab. Αν στα δεδομένα υπάρχουν ελλιπείς τιμές τότε αυτές μαρκάρονται με “ ? ”.

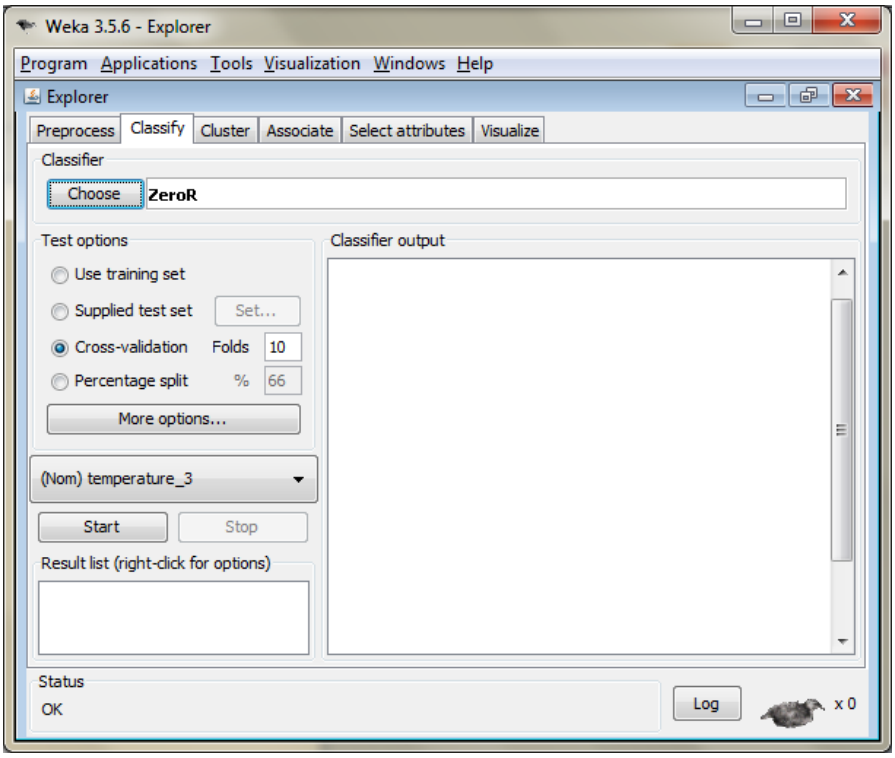
```
@relation weka_validation_set_1

@attribute temperature_1 numeric
@attribute humidity_1 numeric
@attribute temperature_2 numeric
@attribute humidity_2 numeric
@attribute humidity_3 numeric
@attribute Wind_speed numeric
@attribute temperature_3 {a,b,c,d}

@data

7.2   103.8 7.7   101.9 104.5 0.02  ?
7.2   103.7 7.8   101.7 104.5 0     ?
7.6   103.7 8.2   101.6 104.5 0     ?
7.7   103.7 8.2   101.6 104.5 0     ?
8     103.6 8.6   101.5 104.5 0.01  ?
8.1   103.6 8.6   101.4 104.5 0.32  ?
8.3   103.4 9     101.2 104.5 0.16  ?
8.9   103.1 9.4   100   104.5 0.28  ?
```

Εικόνα 13: Format αρχείου ARFF

Αφού εισάγουμε τα δεδομένα (Εικόνα 14) μπορούμε να ξεκινήσουμε την επεξεργασία τους. Όλοι οι αλγόριθμοι που περιγράφηκαν σε αυτό το κεφάλαιο περιλαμβάνονται στο Weka, στην καρτέλα (Tab) “Classify”. Εκεί μπορούμε να επιλέξουμε ποιον αλγόριθμο θέλουμε να χρησιμοποιήσουμε πατώντας το κουμπί “Choose”. (Εικόνα 15)



Εικόνα 14: Εισαγωγή Δεδομένων στο πρόγραμμα WEKA

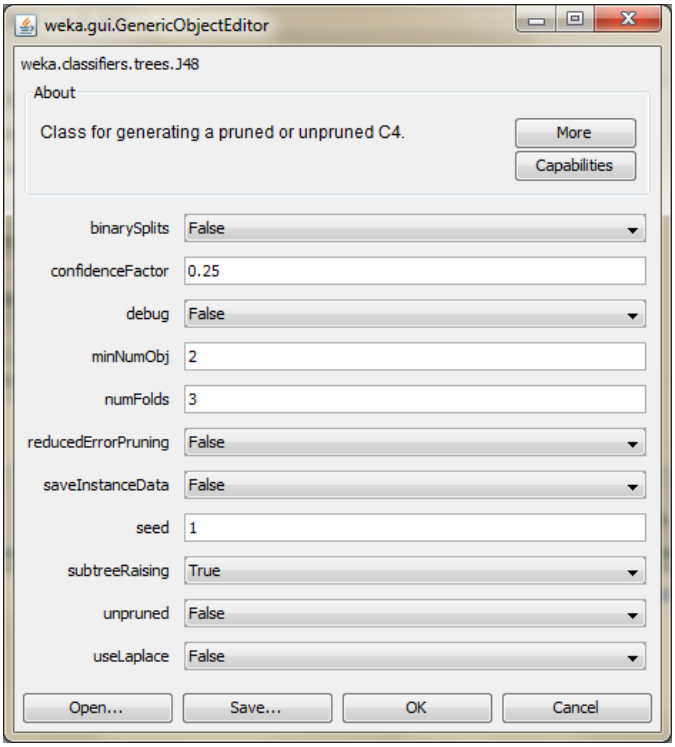


Εικόνα 15: Επιλογή Ταξινομητή (Classifier) στο πρόγραμμα WEKA

Ένας από τους διαθέσιμους classifiers είναι ο J48 ο οποίος υλοποιεί τον αλγόριθμο C4.5. Αφού τον επιλέξουμε μπορούμε να κάνουμε κλικ πάνω στην γραμμή που αναγράφεται και να εμφανιστεί το παράθυρο επεξεργασίας του αλγορίθμου (Εικόνα 16). Από τις επιλογές που υπάρχουν, μπορούμε να επιλέξουμε εάν το δένδρο που θα κατασκευαστεί θα είναι δυαδικό ή όχι. Επίσης, μπορούμε να ρυθμίσουμε το όριο

εμπιστοσύνης για την διαδικασία του pruning (προεπιλογή 0,25) και να επιλέξουμε τον ελάχιστο αριθμό αντικειμένων που θα περιέχει το κάθε φύλλο (προεπιλογή 2). Αντί για το κανονικό pruning μπορούμε να επιλέξουμε το reduced-error pruning. Η παράμετρος numFolds (προεπιλογή 3) καθορίζει το μέγεθος του pruning set, δηλαδή τα δεδομένα κατανέμονται ισομερώς σε αυτόν τον αριθμό των τμημάτων που έχει οριστεί και το τελευταίο τμήμα χρησιμοποιείται για pruning. Εάν ορίσουμε την μεταβλητή saveInstanceData “True”, τότε κατά την απεικόνιση του δένδρου θα εμφανίζονται και τα σημεία των αρχικών δεδομένων (αυτή η μεταβλητή έχει οριστεί “False” ως προεπιλογή για ελαχιστοποίηση των απαιτήσεων της μνήμης) [26].

Οι υπόλοιποι αλγόριθμοι που περιγράφηκαν σε αυτό το κεφάλαιο έχουν αντίστοιχες επιλογές για την αλλαγή των παραμέτρων τους.



Εικόνα 16: Παράμετροι του αλγορίθμου J48

Ο αλγόριθμος DecisionStump κατασκευάζει ένα μοντέλο δυαδικού δένδρου, 1 - επιπέδου για τα αριθμητικά, δεδομένα εισόδου εκμάθησης. Τις ελλιπείς τιμές τις αντιμετωπίζει σαν μια νέα παράμετρο και δημιουργεί ένα ξεχωριστό φύλλο, ορίζοντας για αυτές μια νέα τιμή.

Ο αλγόριθμος M5P δημιουργεί ένα δένδρο όπου κάνει πρόβλεψη για την μεταβλητή που ορίσαμε και στα φύλλα του δένδρου αποθηκεύει μια γραμμική συνάρτηση με παραμέτρους τις υπόλοιπες μεταβλητές του συνόλου των δεδομένων.

Αντίστοιχα και ο αλγόριθμος RepTree κατασκευάζει ένα δένδρο με τη διαφορά ότι στα φύλλα αποθηκεύει σταθερές τιμές.

Αφού δημιουργηθούν τα μοντέλα των δένδρων από τα εκπαιδευτικά δεδομένα (training set), μετατρέπουμε σε java κώδικα τα μοντέλα και δίνουμε ως είσοδο το υπόλοιπο σύνολο δεδομένων, το οποίο περιλαμβάνει και ελλιπείς τιμές. Ο κώδικας, επιστέφει την προβλεπόμενη τιμή βάση του αντίστοιχου μοντέλου για κάθε ελλιπή τιμή

και με την κατάλληλη επεξεργασία των δεδομένων εξόδου μπορούμε να εξάγουμε τα σχετικά συμπεράσματα.

Τέλος, οι βιβλιοθήκες του WEKA διατίθενται και σε jar αρχεία ώστε να μπορούν να χρησιμοποιηθούν απευθείας σε java προγραμματισμό.

4.3.2 Matlab

Το εργαλείο Matlab 7.11.0 (R2010b) χρησιμοποιήθηκε για την υλοποίηση ενός δένδρου απόφασης για πρόβλεψη τιμών [28]. Η συνάρτηση ***classregtree(X,y,'Name',value)*** δημιουργεί ένα δένδρο απόφασης για να προβλέψει τις τιμές της μεταβλητής y σε συνάρτηση των τιμών που περιλαμβάνονται στον πίνακα X. Ο X είναι ένας πίνακας n x m που περιέχει τις τιμές των μετρήσεων. Το y είναι ένα διάνυσμα μεγέθους n όπου έχουν αποθηκευτεί οι τιμές της μεταβλητής που θα χρειαστεί μελλοντικά να προβλεφθούν οι τιμές της. Δηλαδή, αφού κατασκευάσουμε το μοντέλο με τα εκπαιδευτικά δεδομένα, θα γίνει πρόβλεψη για τις ελλιπείς τιμές που θα παρουσιαστούν στην μεταβλητή y. Τα ορίσματα της συνάρτησης classregtree: “Name” και “value” χρησιμοποιούνται για να δώσουν τιμές στις στήλες του πίνακα X, δηλαδή στις μεταβλητές των δεδομένων εισόδου.

Με την χρήση των εντολών που παρέχονται από το Matlab, μπορούμε να υπολογίσουμε το Cross-validation error και το Resubstitution error. Το Resubstitution error είναι το ποσοστό των μετρήσεων του συνόλου εκπαίδευσης που ταξινομούνται λάθος με βάση το μοντέλο που έχει κατασκευαστεί. Για την εύρεση του Cross-validation error, το training dataset χωρίζεται σε υποσύνολα περίπου ίδιου μεγέθους και ίδιας αναλογίας αντικειμένων σε κάθε υποσύνολο. Από αυτά τα υποσύνολα, αφαιρείται το ένα και τα υπόλοιπα χρησιμοποιούνται ως training set για να δημιουργηθεί ένα classification model. Τέλος, γίνεται ταξινόμηση του υποσυνόλου που αφαιρέθηκε και έτσι υπολογίζεται το cross-validation error. Αν το cross-validation error είναι κατά πολύ μεγαλύτερο από το resubstitution error, τότε το παραγόμενο δένδρο υπερκαλύπτει το training set. Με άλλα λόγια, το δένδρο ταξινομεί πολύ καλά το αρχικό σύνολο δεδομένων αλλά η δομή του είναι αρκετά εύθικτη και προσαρμοσμένη σε αυτό το σύνολο δεδομένων με αποτέλεσμα να υποβαθμίζεται η αξιοπιστία του σε ένα νέο σύνολο δεδομένων. Σε αυτές τις περιπτώσεις είναι προτιμότερο να βρεθεί ένα απλούστερο δένδρο που θα αποδίδει καλύτερα στα νέα δεδομένα απ΄ότι ένα σύνθετο δένδρο.

Για την απλοποίηση του δένδρου χρησιμοποιήθηκε η εντολή ***prune(t,bestlevel)***. Η εντολή prune, παίρνει το αρχικό δένδρο t και το “κλαδεύει” στο επίπεδο που ορίζεται στην μεταβλητή bestlevel. Για να ελέγξουμε ποιο είναι το καλύτερο επίπεδο κλαδέματος του δένδρου, υπολογίζουμε μια τιμή cutoff η οποία ισούται με το ελάχιστο κόστος συν ένα τυπικό σφάλμα. Το bestlevel που υπολογίζεται από την μέθοδο classregtree είναι το μικρότερο δένδρο κάτω από το cutoff. Όταν η μεταβλητή bestlevel = 0 τότε αντιστοιχεί σε “ακλάδευτο” δέντρο, για αυτό το λόγο αν θέλουμε να την χρησιμοποιήσουμε ως δείκτης σε διάνυσμα πρέπει να προσθέσουμε 1.

Για την οπτικοποίηση του δένδρου υπάρχει η εντολή ***view(t)***. Αφού δημιουργηθεί το δένδρο που θα χρησιμοποιηθεί ως μοντέλο πρόβλεψης, εκτελούμε την εντολή ***eval(t, data)***. Η μέθοδος eval κάνει πρόβλεψη των τιμών για την μεταβλητή y με βάση τις τιμές που υπάρχουν στον πίνακα data. Ο πίνακας data πρέπει να έχει ίδια δομή με το πίνακα X, αφού βάσει αυτού δημιουργήθηκε το μοντέλο (classification tree).

5 ΠΕΙΡΑΜΑΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ ΣΥΣΤΗΜΑΤΟΣ ΕΚΤΙΜΗΣΗΣ ΕΛΛΙΠΟΥΣ ΠΛΗΡΟΦΟΡΙΑΣ

5.1 Δεδομένα που χρησιμοποιήθηκαν (Datasets)

Για την εφαρμογή των πειραμάτων, χρησιμοποιήθηκαν δεδομένα που συλλέχθηκαν σε πραγματικό χρόνο από το πεδίο. Το δίκτυο αισθητήρων που χρησιμοποιήθηκε αποτελείται από τρία ζευγάρια αισθητήρων που μετρούν την θερμοκρασία και την υγρασία του περιβάλλοντος και έναν τέταρτο αισθητήρα που μετρά την ταχύτητα του ανέμου τις ίδιες χρονικές στιγμές. Συνεπώς, η κάθε μία διανυσματική μέτρηση αποτελείται από επτά μεταβλητές της μορφής:

x = (temp_1, hum_1, temp_2, hum_2, temp_3, hum_3, wind_speed)

Οι τιμές temp_1, hum_1 προέρχονται από την μέτρηση της θερμοκρασίας και της υγρασίας αντίστοιχα του πρώτου ζευγαριού αισθητήρων. Ομοίως, οι τιμές temp_2, hum_2, temp_3, hum_3 προέρχονται από το δεύτερο και τρίτο αισθητήριο κόμβο, αντίστοιχα. Τέλος, η τιμή wind_speed προέρχεται από τις μετρήσεις του τέταρτου αισθητήρα.

Τα δεδομένα που έχουν συλλεχθεί και χρησιμοποιηθεί στα πειράματα απεικονίζουν τις μετρήσεις για 387 διαφορετικές χρονικές στιγμές. Συνεπώς, έχουμε 387 διαφορετικές μετρήσεις για διακριτές χρονικές στιγμές.

Τα πειράματα που έγιναν καθώς και τα συμπεράσματα που εξήχθησαν, μπορούν να επεκταθούν και σε ένα γενικότερο περιβάλλον με πλήθος κόμβων αισθητήρων (άρα και διαφορετικό αριθμό μεταβλητών) καθώς και σε διαφορετικό είδος μετρήσεων.

5.2 Πειράματα Προεκβολής

5.2.1 Σενάρια που δοκιμάστηκαν

Για τον έλεγχο της ακρίβειας μιας πρόβλεψης που βασίζεται στην προεκβολή δε λάβαμε υπόψιν όλες τις τιμές από την αρχή του dataset αλλά μόνο τις m τελευταίες μετρήσεις για να προβλέψουμε την αμέσως επόμενη τιμή του κάθε γνωρίσματος. Το μέγεθος του m ξεκίνησε από 2 και έφτασε μέχρι 10. Έπειτα από αρκετά πειράματα παρατηρήθηκε ότι όταν το m έπαιρνε τιμές μεγαλύτερες του 10, οι τιμές που προέβλεπε συνέκλιναν οπότε δεν υπήρχε λόγος να εξετάσουμε προβλέψεις για μεγαλύτερη τιμή m. Για κάθε τιμή του m, υπολογίσαμε την τιμή πρόβλεψης που προέκυπτε έπειτα από την εφαρμογή της μεθοδολογίας Cubic Spline καθώς και την μέση τιμή του γνωρίσματος βάσει των προηγούμενων μετρήσεων.

Πιο συγκεκριμένα, σε κάθε σενάριο δόθηκαν διαδοχικά m μετρήσεις με σκοπό να δημιουργηθεί η αντίστοιχη συνάρτηση και έτσι να γίνει η πρόβλεψη για την $251^{η}$ χρονική στιγμή. Πραγματοποιήθηκαν προβλέψεις για όλες τις μεταβλητές ξεχωριστά. Δηλαδή στο πρώτο σενάριο χρησιμοποιήθηκαν τα δεδομένα για να προβλέψουν την τιμή της θερμοκρασίας από τον πρώτο αισθητήρα (temp_1), στο δεύτερο σενάριο την τιμή της υγρασίας από το δεύτερο αισθητήρα (hum_1) κ.ο.κ. Με αυτό τον τρόπο προκύπτουν οι

προβλέψεις για όλες τις μεταβλητές για την 251η χρονική στιγμή και για όλα τα μεγέθη του m.

Ο σκοπός που δοκιμάστηκαν σενάρια με διαφορετικό πλήθος μετρήσεων είναι για να δούμε πώς το μέγεθος του παραθύρου (m) επηρεάζει την ακρίβεια της συνάρτησης που κατασκευάζεται και συνεπώς την ακρίβεια των αποτελεσμάτων. Επίσης θέλαμε να συγκρίνουμε τα αποτελέσματα που προκύπτουν από την εφαρμογή της Cubic Spline με αυτά της μέσης τιμής της ίδιας της μεταβλητής ώστε να ελέγξουμε ποιά μέθοδος είναι πιο αποτελεσματική και αξιόπιστη.

5.2.2 Παραδείγματα εκτίμησης τιμών

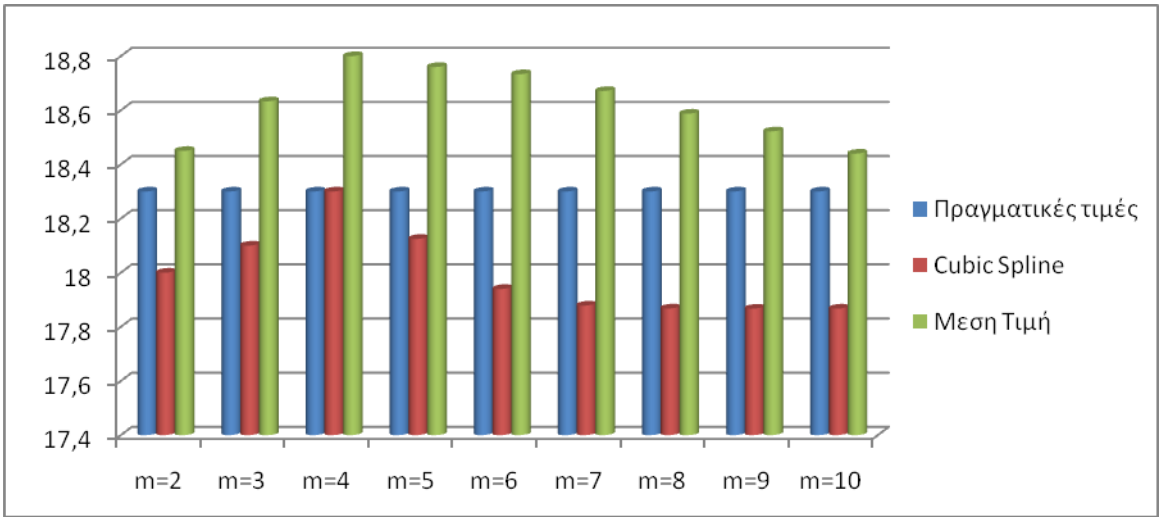
1ο Σενάριο – Πρόβλεψη Θερμοκρασίας 1ου Ζεύγους Αισθητήρων

Στην πρώτη στήλη αναγράφεται το πλήθος των τελευταίων δεδομένων που χρησιμοποιήθηκαν ως είσοδο για την πρόβλεψη της 251ης χρονικής στιγμής. Αφού κάθε φορά η συνάρτηση κατασκευάζεται από m μετρήσεις, δηλαδή από μετρήσεις m διαδοχικών χρονικών στιγμών, η πρόβλεψη γίνεται για την (m+1)η χρονική στιγμή. Στην δεύτερη στήλη αναγράφονται οι τιμές που προέκυψαν από την εφαρμογή της Cubic Spline ενώ στη τρίτη στήλη είναι η μέση τιμή των m προηγούμενων μετρήσεων της θερμοκρασίας του πρώτου αισθητήρα. Να σημειωθεί ότι η πραγματική μέτρηση που έχει παρθεί από τον ίδιο τον αισθητήρα είναι ***18,3***.

Πίνακας 1: Αποτελέσματα Πρόβλεψης 1ης Θερμοκρασίας

Μέγεθος παραθύρου	Αποτέλεσμα Cubic Spline	Αποτέλεσμα Μέσης Τιμής
m=2	18,0000	18,4500
m=3	18,1000	18,6333
m=4	18,3000	18,8000
m=5	18,1250	18,7600
m=6	17,9400	18,7333
m=7	17,8786	18,6714
m=8	17,8679	18,5875
m=9	17,8672	18,5222
m=10	17,8674	18,4400

Στο Σχήμα 1 βλέπουμε την διαγραμματική απεικόνιση των αποτελεσμάτων του πρώτου σεναρίου. Όπως παρατηρούμε, όσο το m παραμένει μικρό οι προβλέψεις της Cubic Spline είναι αρκετά ακριβείς. Πιο συγκεκριμένα για m=4 πρόβλεψη έχει 100% ακρίβεια και μέχρι το m = 6 η Cubic Spline δίνει πλησιέστερα αποτελέσματα στην πραγματική μέτρηση απ'ότι η μέση τιμή. Ωστόσο και η μέση τιμή δίνει αρκετά ικανοποιητικά αποτελέσματα κυρίως καθώς το m αυξάνεται. Δηλαδή για m > 6 είναι προτιμότερο να χρησιμοποιήσουμε την πρόβλεψη της μέσης τιμής φαίνεται ότι προσεγγίζει καλύτερα την πραγματική τιμή από την Cubic Spline που σταδιακά σταθεροποιείται περίπου στο 17,86.



Σχήμα 1: Διαγραμματική Απεικόνιση Πρόβλεψης 1ης Θερμοκρασίας

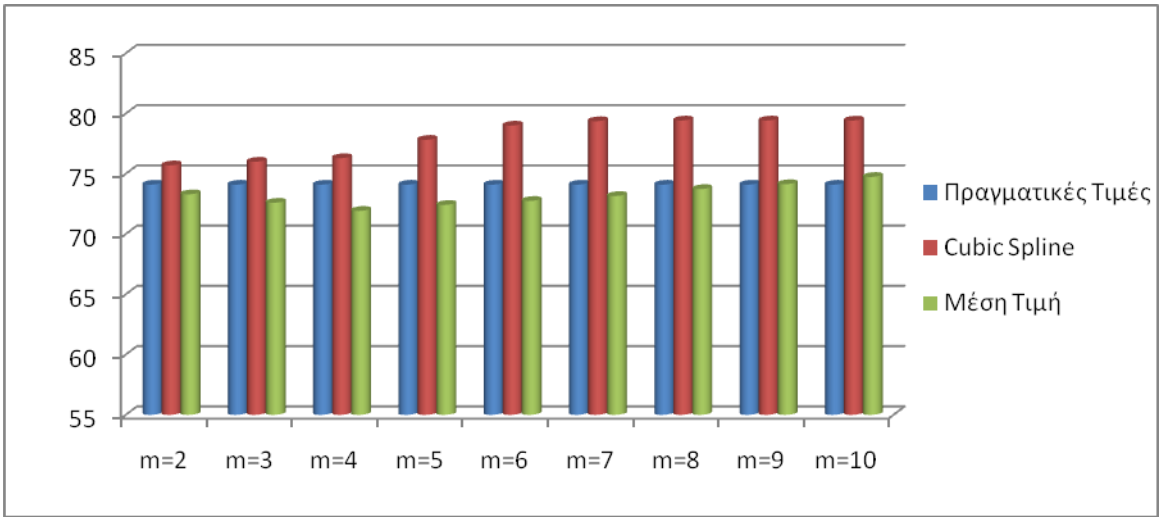
2ο Σενάριο - Πρόβλεψη Υγρασίας 1ου Ζεύγους Αισθητήρων

Σε αυτό το σενάριο παραθέτονται οι προβλέψεις που έγιναν για την υγρασία του πρώτου ζεύγους αισθητήρων την 251η χρονική στιγμή. Η πραγματική τιμή της μέτρηση είναι ***74,1***.

Πίνακας 2: Αποτελέσματα Πρόβλεψης 1ης Υγρασίας

Μέγεθος παραθύρου	Αποτέλεσμα Cubic Spline	Αποτέλεσμα Μέσης Τιμής
m=2	75,7000	73,3000
m=3	76,0000	72,6000
m=4	76,3000	71,925
m=5	77,8250	72,4200
m=6	79,0000	72,7500
m=7	79,3696	73,1571
m=8	79,4244	73,7375
m=9	79,4240	74,1444
m=10	79,4207	74,7400

Στο δεύτερο σενάριο τα αποτελέσματα των προβλέψεων είναι καλύτερα με τη χρήση της μέσης τιμής απ'ότι με τη Cubic Spline, για όλα τα μεγέθη του m. Παρατηρούμε ότι καθώς αυξάνεται το m, η μέση τιμή βελτιώνει την απόδοσή της ενώ το αντίθετο συμβαίνει με τη Cubic Spline.



Σχήμα 2: Διαγραμματική Απεικόνιση Πρόβλεψης 1ης Υγρασίας

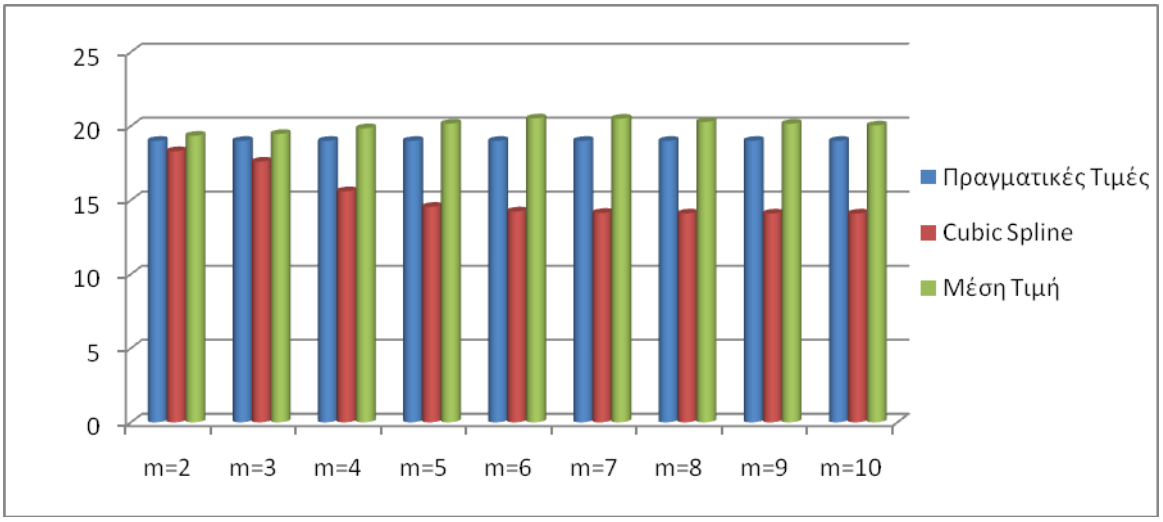
3ο Σενάριο - Πρόβλεψη Θερμοκρασίας 2ου Ζεύγους Αισθητήρων

Σε αυτό το σενάριο παραθέτονται οι προβλέψεις που έγιναν για την θερμοκρασία του δεύτερου ζεύγους αισθητήρων την 251η χρονική στιγμή. Η πραγματική τιμή της μέτρηση είναι ***19***.

Πίνακας 3: Αποτελέσματα Πρόβλεψης 2ης Θερμοκρασίας

Μέγεθος παραθύρου	Αποτέλεσμα Cubic Spline	Αποτέλεσμα Μέσης Τιμής
m=2	18,3000	19,3500
m=3	17,6000	19,4666
m=4	15,6000	19,8500
m=5	14,5500	20,1600
m=6	14,2400	20,5166
m=7	14,1339	20,5000
m=8	14,0957	20,2875
m=9	14,0910	20,1555
m=10	14,0920	20,0400

Στο τρίτο σενάριο τα αποτελέσματα των προβλέψεων είναι ανάλογα με το δεύτερο σενάριο. Η πρόβλεψη που γίνεται με βάση την μέση τιμή είναι καλύτερη απ'ότι με τη Cubic Spline, για όλα τα μεγέθη του m.



Σχήμα 3: Διαγραμματική Απεικόνιση Πρόβλεψης 2ης Θερμοκρασίας

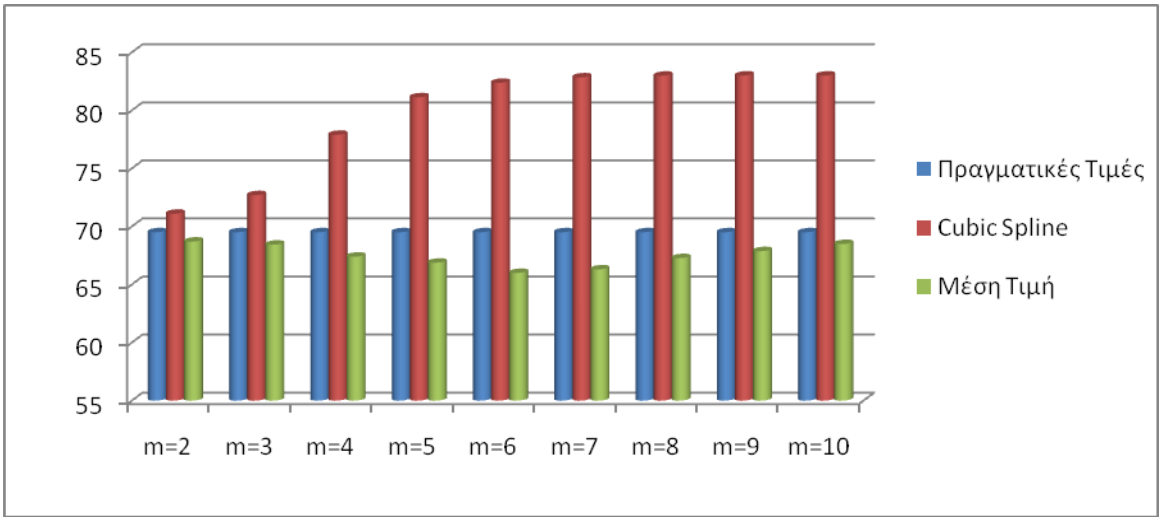
4ο Σενάριο - Πρόβλεψη Υγρασίας 2ου Ζεύγους Αισθητήρων

Σε αυτό το σενάριο παραθέτονται οι προβλέψεις που έγιναν για την υγρασία του δεύτερου ζεύγους αισθητήρων την 251η χρονική στιγμή. Η πραγματική τιμή της μέτρηση είναι ***69,5***.

Πίνακας 4: Αποτελέσματα Πρόβλεψης 2ης Υγρασίας

Μέγεθος παραθύρου	Αποτέλεσμα Cubic Spline	Αποτέλεσμα Μέσης Τιμής
m=2	71,1000	68,7000
m=3	72,7000	68,4333
m=4	77,9000	67,4000
m=5	81,1250	66,8800
m=6	82,3733	66,0000
m=7	82,8411	66,3000
m=8	82,9876	67,2750
m=9	83,0023	67,8777
m=10	82,9975	68,4900

Και στο τέταρτο σενάριο τα αποτελέσματα των προβλέψεων είναι καλύτερα με τη χρήση της μέσης τιμής απ'ότι με τη Cubic Spline, για όλα τα μεγέθη του m. Παρατηρούμε ότι καθώς αυξάνεται το m, οι προβλέψεις της Cubic Spline αποκλίνουν από την πραγματική τιμή ενώ η μέση τιμή κυμαίνεται σε πιο κοντινά επίπεδα με την πραγματική μέτρηση.



Σχήμα 4: Διαγραμματική Απεικόνιση Πρόβλεψης 2ης Υγρασίας

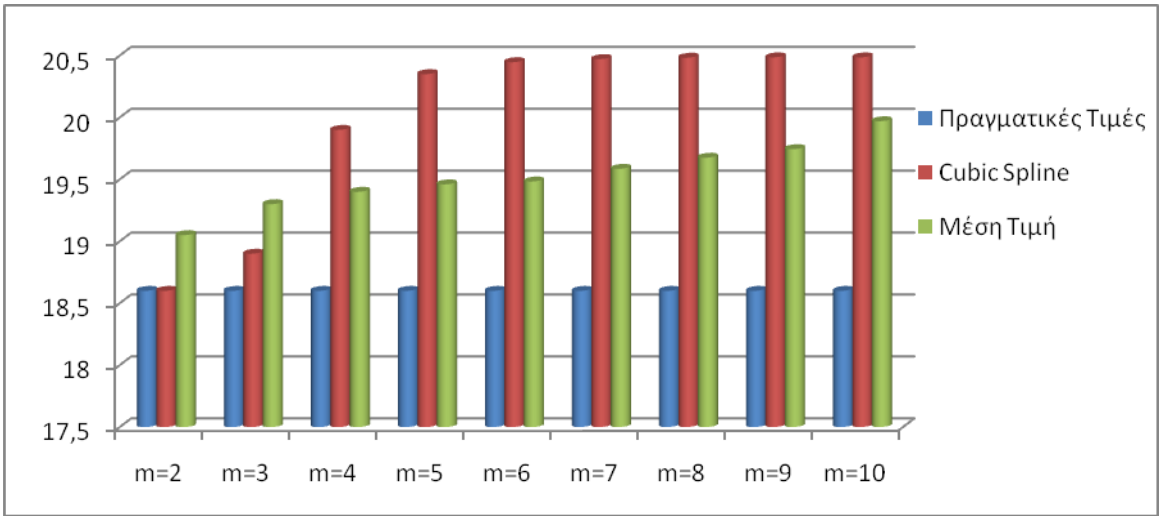
5ο Σενάριο - Πρόβλεψη Θερμοκρασίας 3ου Ζεύγους Αισθητήρων

Σε αυτό το σενάριο παραθέτονται οι προβλέψεις που έγιναν για την θερμοκρασία του τρίτου ζεύγους αισθητήρων την 251η χρονική στιγμή. Η πραγματική τιμή της μέτρηση είναι ***18,6***.

Πίνακας 5: Αποτελέσματα Πρόβλεψης 3ης Θερμοκρασίας

Μέγεθος παραθύρου	Αποτέλεσμα Cubic Spline	Αποτέλεσμα Μέσης Τιμής
m=2	18,6000	19,0500
m=3	18,9000	19,3000
m=4	19,9000	19,4000
m=5	20,3500	19,4600
m=6	20,4467	19,4833
m=7	20,4714	19,5857
m=8	20,4828	19,6750
m=9	20,4856	19,7444
m=10	20,4854	19,9700

Και σε αυτό το σενάριο τα αποτελέσματα των προβλέψεων είναι τα αναμενόμενα. Οι προβλέψεις που γίνονται με βάση την μέση τιμή είναι καλύτερες απ'ότι με τη Cubic Spline, για τα περισσότερα μεγέθη του m. Μόνο για m <=3 υπερτερεί η Cubic Spline.



Σχήμα 5: Διαγραμματική Απεικόνιση Πρόβλεψης 3ης Θερμοκρασίας

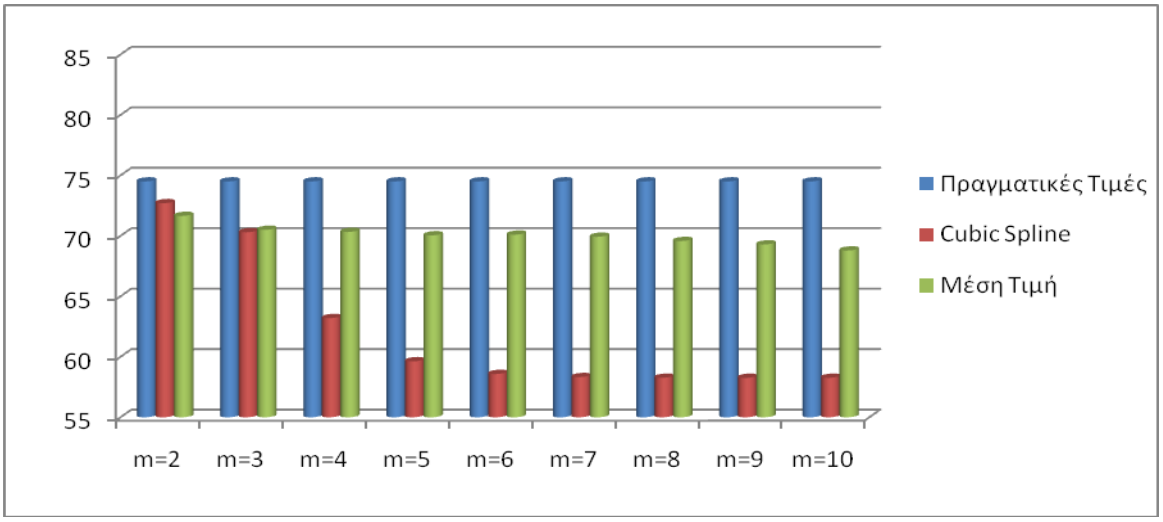
6ο Σενάριο - Πρόβλεψη Υγρασίας 3ου Ζεύγους Αισθητήρων

Σε αυτό το σενάριο παραθέτονται οι προβλέψεις που έγιναν για την υγρασία του τρίτου ζεύγους αισθητήρων την 251η χρονική στιγμή. Η πραγματική τιμή της μέτρηση είναι ***74,5***.

Πίνακας 6: Αποτελέσματα Πρόβλεψης 3ης Υγρασίας

Μέγεθος παραθύρου	Αποτέλεσμα Cubic Spline	Αποτέλεσμα Μέσης Τιμής
m=2	72,7000	71,6500
m=3	70,3000	70,5000
m=4	63,2000	70,3250
m=5	59,6250	70,0400
m=6	58,5800	70,1000
m=7	58,3214	69,9285
m=8	58,2641	69,5750
m=9	58,2591	69,2880
m=10	58,2609	68,7900

Αντίστοιχα με το 5ο σενάριο, επιβεβαιώνεται και εδώ, ότι μόνο για πολύ μικρό m η Cubic Spline δίνει καλύτερα αποτελέσματα. Παρότι και η μέση τιμή αποκλίνει καθώς αυξάνεται το m, φαίνεται να διατηρεί καλύτερες τιμές. Ο λόγος που και οι δύο μεθοδολογίες μειώνουν τις τιμές των προβλέψεων τους καθώς αυξάνεται το m (αν και η πραγματική τιμή είναι μεγαλύτερη), οφείλεται καθαρά στο δείγμα των δεδομένων εισόδου που περιέχει χαμηλότερες μετρήσεις.



Σχήμα 6: Διαγραμματική Απεικόνιση Πρόβλεψης 3ης Υγρασίας

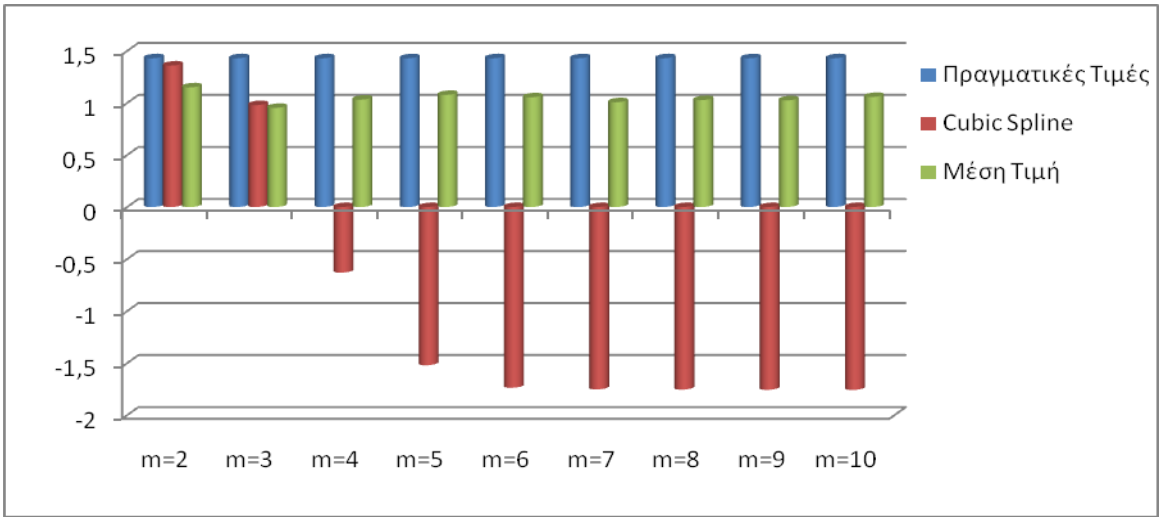
7ο Σενάριο Πρόβλεψη Ταχύτητας Ανέμου

Σε αυτό το σενάριο παραθέτονται οι προβλέψεις που έγιναν για την ταχύτητα του ανέμου την 251η χρονική στιγμή. Η πραγματική τιμή της μέτρηση είναι ***1,43***.

Πίνακας 7: Αποτελέσματα 3ου Σεναρίου Προεκβολής

Μέγεθος παραθύρου	Αποτέλεσμα Cubic Spline	Αποτέλεσμα Μέσης Τιμής
M=2	1,3600	1,1500
M=3	0,9800	0,9533
M=4	-0,6300	1,0325
M=5	-1,5200	1,0780
M=6	-1,7367	1,0550
M=7	-1,7521	1,0071
M=8	-1,7540	1,0275
M=9	-1,7565	1,0255
m=10	-1,7575	1,0580

Στα αποτελέσματα των προβλέψεων παρατηρούμε μεγάλη αστοχία από την μεθοδολογία Cubic Spline. Το μεγαλύτερο μέρος των προβλέψεων της αντιστοιχεί σε αρνητικές τιμές, πράγμα που δε μπορεί να ισχύει σε καμία περίπτωση. Η εξήγηση είναι πως οι μετρήσεις που δίνονται ως είσοδο έχουν μεγάλες αποκλίσεις μεταξύ τους και σε πολλά σημεία παρουσιάζουν κατακόρυφη πτώση των τιμών τους. Συνεπώς, η καμπύλη που δημιουργείται από την Cubic Spline, θεωρεί πως οι τιμές θα συνεχίσουν να μειώνονται και στις επόμενες χρονικές στιγμές και πέφτει κάτω από το μηδέν. Δεν λαμβάνει υπόψιν το γεγονός ότι όλες οι τιμές εισόδου είναι θετικές ώστε να αποκλείσει την πρόβλεψη αρνητικής τιμής.



Σχήμα 7: Διαγραμματική Απεικόνιση Ταχύτητας Ανέμου

5.3 Πειράματα Κατηγοριοποίησης

5.3.1 Σενάρια που δοκιμάστηκαν

Για τον έλεγχο των αλγορίθμων κατηγοριοποίησης, χρησιμοποιήθηκε το dataset που περιγράφθηκε στην ενότητα 5.1. Στα πειράματα που εφαρμόστηκαν, υλοποιήθηκαν τρία σενάρια. Σε κάθε ένα σενάριο δόθηκε ως είσοδο ένα διαφορετικό ποσοστό του αρχικού δείγματος με στόχο να προβλέψουν ένα τμήμα από τις αρχικές μετρήσεις. Το πρώτο σενάριο χρησιμοποίησε περίπου το 1/20 του συνόλου ως δεδομένα εκπαίδευσης (20 μετρήσεις). Το δεύτερο σενάριο χρησιμοποίησε περίπου το 1/5 του συνόλου ως δεδομένα εκπαίδευσης (75 μετρήσεις) ενώ το τρίτο σενάριο το ½ του συνόλου (190 μετρήσεις). Για κάθε σενάριο, το κομμάτι των δεδομένων που θα χρησιμοποιούταν για επαλήθευση, παρέμενε σταθερό και ήταν 50 μετρήσεις. Δηλαδή και στα τρία σενάρια τα δεδομένα επαλήθευσης παρέμεναν σταθερά και κάθε φορά αυξανόταν το σύνολο των δεδομένων εκπαίδευσης. Ουσιαστικά το σύνολο εκπαίδευσης είναι ένα παράθυρο που κάθε φορά αυξάνεται για να προβλέψει τις επόμενες 50μετρήσεις. Ο λόγος είναι για να δούμε τι επίδραση θα έχει αυτό στο δένδρο που δημιουργείται και συνεπώς στις προβλέψεις.

Για τα δεδομένα που χρησιμοποιήθηκαν στην επαλήθευση των προβλέψεων, επιλέχθηκε μια μεταβλητή μέτρησης για την οποία θα γίνουν οι προβλέψεις. Για τα πειράματα της παρούσας διπλωματικής εργασίας, επιλέχθηκαν οι μετρήσεις της θερμοκρασίας του τρίτου ζεύγους αισθητήρων (δηλαδή η πέμπτη παράμετρος των μετρήσεων). Θεωρήσαμε ότι οι τιμές αυτών των μετρήσεων δεν είχαν συλλεχθεί, οπότε αφαιρέθηκαν από τον πίνακα. Τέλος, έγινε η πρόβλεψη των ελλιπών τιμών με βάση τα δένδρα που είχαν δημιουργηθεί από τα δεδομένα εκπαίδευσης και με είσοδο στους αλγορίθμους τις υπόλοιπες μετρήσεις των αισθητήρων.

Ορισμένοι από τους αλγόριθμους που χρησιμοποιήθηκαν στα πειράματα δεν εφαρμόζονται σε συνεχή δεδομένα (π.χ. ο αλγόριθμος C4.5 εφαρμόζεται μόνο σε διακριτά δεδομένα). Συνεπώς έπρεπε να γίνει μια προεπεξεργασία των δεδομένων εισόδου ώστε να μετατραπούν σε διακριτές οι τιμές της στήλης για πρόβλεψη (στην περίπτωσή μας θερμοκρασία $3^{ου}$ αισθητήρα, δηλαδή $5^{η}$ στήλη). Βέβαια, υπάρχει και η περίπτωση κάποιων άλλων αλγορίθμων (όπως ο M5P) που εφαρμόζονται μόνο σε συνεχείς τιμές και όχι σε διακριτές. Σε αυτή την περίπτωση χρησιμοποιήθηκε το αρχικό

dataset που περιείχε τις συνεχείς τιμές και αφού έγινε η πρόβλεψη, έπειτα έγινε η μετατροπή σε διακριτές τιμές ώστε να γίνει η σύγκριση των αποτελεσμάτων. Το πλήθος των διακριτών τιμών που θα αντικαθιστούσαν τις πραγματικές τιμές (συνεχείς τιμές) επιλέχθηκε τυχαία. Άρα οι τιμές τις θερμοκρασίας αντικαταστάθηκαν με τις εξής διακριτές τιμές:

a : περιλαμβάνει θερμοκρασίες του εύρους (6 - 9,9) βαθμούς °C.

b : περιλαμβάνει θερμοκρασίες του εύρους (10 - 13,9) βαθμούς °C.

c : περιλαμβάνει θερμοκρασίες του εύρους (14 - 17,9) βαθμούς °C.

d : περιλαμβάνει θερμοκρασίες του εύρους (18 - 22,3) βαθμούς °C.

Σε δεύτερη φάση, τα παραπάνω σενάρια εκτελέστηκαν εκ νέου, αφού πρώτα εφαρμόστηκε η τεχνική PCA. Σκοπός ήταν να δούμε την απόκλιση των προβλέψεων όταν εφαρμόζονται τεχνικές μείωσης διαστάσεων. Για να μπορέσουμε να συγκρίνουμε την απόκλιση ανάμεσα στην τιμή πρόβλεψης του αρχικού μοντέλου και στην πρόβλεψη έπειτα από την εφαρμογή PCA, συγκρίναμε τα αποτελέσματα από τους αλγορίθμους που προβλέπουν πραγματικές τιμές (δηλαδή M5P, Decision Stump και RepTree). Σε ένα πραγματικό περιβάλλον, όπου θα υπάρχει συνεχής ροή μετρήσεων ως είσοδο, είναι αναπόφευκτη η χρήση τεχνικών μείωσης διαστάσεων. Συνεπώς αυτό επηρεάζει την απόδοση των αλγορίθμων του εξετάζουμε.

Επίσης, όλα τα σενάρια εφαρμόστηκαν σε δύο διαφορετικά σύνολα δεδομένων (εκπαίδευσης και επαλήθευσης) για να εξετάσουμε αν η απόδοση των αλγορίθμων παραμένει σταθερή ή επηρεάζεται από το είδος των μετρήσεων.

5.3.2 Μετρικές

Βασικός στόχος των πειραμάτων είναι να μετρηθεί η ακρίβεια των προβλέψεων και η απώλεια πληροφορίας έπειτα από την εφαρμογή PCA. Για να επιτευχθεί αυτό, συγκρίνουμε τα αποτελέσματα της πρόβλεψης με τις πραγματικές τιμές και υπολογίζουμε το ποσοστό επιτυχίας. Ύστερα, εφαρμόζουμε PCA και υπολογίζουμε το σχετικό σφάλμα και την διακύμανση σε σχέση με την αρχική πρόβλεψη.

Στα πειράματα που έγιναν με χρήση του Decision Tree (Matlab), δημιουργήσαμε αρχικά το Decision Tree για να μετρήσουμε το ποσοστό επιτυχίας. Αυτό το αρχικό δένδρο κατηγοριοποιεί σωστά το training set, αλλά η δομή του δέντρου είναι αρκετά συσχετισμένη με το σύνολο των μετρήσεων που έχει ως αποτέλεσμα την μείωση της απόδοσής του όταν εισάγονται νέα δεδομένα. Τις περισσότερες φορές είναι προτιμότερο ένα απλούστερο δένδρο που θα έχει καλύτερη απόδοση στα νέα δεδομένα απ'ότι ένα πιο σύνθετο δένδρο. Άρα για το ίδιο σύνολο δεδομένων, κατασκευάσαμε και ένα απλούστερο δένδρο (pruned tree) ώστε να γίνει πρόβλεψη τιμών βάσει αυτού και τελικά να συγκριθεί η ακρίβεια των αποτελεσμάτων.

Στα πειράματα των αλγορίθμων C4.5, M5P, Decision Stump και RepTree, που υλοποιήθηκαν με το εργαλείο WEKA, εκτός από το ποσοστό επιτυχίας, καταμετρήθηκε η Τετραγωνική Ρίζα Μέσου Τετραγωνικού Σφάλματος (Root Mean Squared Error) και ο Συντελεστής Συσχέτισης (Correlation Coefficient) . Ο συντελεστής συσχέτισης έχει

υπολογιστεί μόνο για τους αλγορίθμους που χρησιμοποιούν αριθμητικές τιμές (δηλαδή: M5P, Decision Stump, RepTree).

Η Τετραγωνική Ρίζα Μέσου Τετραγωνικού Σφάλματος είναι ένας συνηθισμένος εκτιμητής της διακύμανσης των σφαλμάτων και χρησιμοποιείται πολύ συχνά. Μετρά την απόσταση της τιμής που προβλέψαμε από την πραγματική τιμή. Συνεπώς μπορούμε να δούμε την επίδραση που έχουν στα αποτελέσματα οι ακραίες τιμές (περιπτώσεις όπου η πρόβλεψη σφάλματος είναι μεγάλη). Είναι ένα στατιστικό στοιχείο που ερμηνεύεται εύκολα, δεδομένου ότι έχει την ίδια μονάδα μέτρησης με την μεταβλητή που απεικονίζονται στους άξονες. Ο Συντελεστής Συσχέτισης μετρά την συσχέτιση μεταξύ πραγματικών τιμών και των τιμών πρόβλεψης. Οι τιμές του συντελεστή συσχέτισης κυμαίνονται από 1, για περιπτώσεις με τέλεια συσχέτιση αποτελεσμάτων, έως -1 όταν τα αποτελέσματα είναι τέλεια συσχετισμένα αρνητικά. Όταν ο συντελεστής ισούται με 0 δεν υπάρχει καμία συσχέτιση. Φυσικά, δεν πρέπει να παρουσιάζονται αρνητικές τιμές σε λογικές περιπτώσεις μεθόδων πρόβλεψης.

Οι τύποι για τον υπολογισμό των παραπάνω μέτρων απόδοσης δίνονται στον ακόλουθο πίνακα:

Πίνακας 8: Τυπολόγιο Μέτρων Απόδοσης για αριθμητικές προβλέψεις

Root Mean Squared Error (Τετραγωνική Ρίζα Μέσου Τετραγωνικού Σφάλματος)	$\sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{n}}$
Correlation Coefficient (Συντελεστής Συσχέτισης)	$\frac{S_{PA}}{\sqrt{S_P S_A}}$ όπου: $S_{PA} = \frac{\sum_i (p_i - \overline{p})(a_i - \overline{a})}{n-1}$, $S_P = \frac{\sum_i (p_i - \overline{p})^2}{n-1}$, $S_A = \frac{\sum_i (a_i - \overline{a})^2}{n-1}$

όπου: p – τιμές πρόβλεψης, a – πραγματικές τιμές

5.3.3 Παραδείγματα εκτίμησης τιμών

Decision Tree (Matlab)

Στο πρώτο πείραμα εξετάζουμε τα αποτελέσματα από την κατασκευή Decision Tree με τη χρήση του εργαλείου Matlab.

Πίνακας 9: Αποτελέσματα Decision Tree χωρίς χρήση PCA

Πλήθος Δεδομένων Εκπαίδευσης	Ποσοστό Επιτυχίας	
	Χωρίς Pruning	Με Pruning
1ο Δείγμα	Προβλέψεις Μετρήσεων 251 – 300	
20 (1ο Δένδρο)	54%	54%
75 (2ο Δένδρο)	82%	82%
190 (3ο Δένδρο)	98%	92%
2ο Δείγμα	Προβλέψεις Μετρήσεων 338-387	
20 (1ο Δένδρο)	42%	42%
75 (2ο Δένδρο)	90%	90%
190 (3ο Δένδρο)	92%	92%

Στον παραπάνω πίνακα αναγράφονται τα ποσοστά επιτυχίας που προέκυψαν από την εφαρμογή του Decision Tree στα δύο διαφορετικά δείγματα που πήραμε από το training set. Το πρώτο δένδρο κάθε δείγματος αντιστοιχεί στα αποτελέσματα που προέκυψαν όταν χρησιμοποιήθηκε το 1/20 του αρχικού πλήθους δεδομένων για την κατασκευή του δένδρου (), το δεύτερο δένδρο όταν χρησιμοποιήθηκε το 1/5 ενώ στο τρίτο το ½. Η στήλη «Χωρίς Pruning» μας δείχνει τα αποτελέσματα από το δένδρο πριν την εφαρμογή Pruning ενώ η στήλη «Με Pruning» τα αποτελέσματα έπειτα από την εφαρμογή. Παρατηρούμε ότι το ποσοστό επιτυχίας παραμένει σταθερό είτε εφαρμόσουμε pruning είτε όχι στην πλειονότητα των πειραμάτων. Για το 1ο δένδρο κάθε δείγματος είναι λογικό να μην υπάρχει διαφορά γιατί το αρχικό δένδρο αποτελείται μόνο από δύο φύλλα και συνεπώς το pruning δεν προκαλεί καμία μεταβολή. Το ίδιο ισχύει και για το 2ο δένδρο που πρώτου δείγματος. Το 2ο δένδρο του δεύτερου δείγματος διαφοροποιείται με τη εφαρμογή του pruning, παρ' όλα αυτά δίνει τα ίδια αποτελέσματα. Ο λόγος είναι ότι οι μεταβολές του δένδρου είναι τόσο μικρές που δεν επιφέρουν καμία βελτίωση στα αποτελέσματα.

Όσον αφορά την απόδοση κάθε δένδρου, παρατηρούμε ότι με τη χρήση του 1/20 των δεδομένων ως δεδομένα εκπαίδευσης τα ποσοστά επιτυχίας ήταν αρκετά χαμηλά (54% & 42% αντίστοιχα για κάθε δείγμα). Αυτό οφείλεται στο λεγόμενο underfitting. Δηλαδή το δένδρο δεν έχει κατασκευαστεί με τα απαραίτητα δεδομένα για να καλύψει τα αποτελέσματα από όλες τις δυνατές περιπτώσεις. Η χρήση του 1/5 ως δεδομένα εκπαίδευσης δίνει πολύ ικανοποιητικά αποτελέσματα (82% & 90% αντίστοιχα για κάθε δείγμα) και τέλος το ½ αυξάνει ακόμα περισσότερο τα ποσοστά επιτυχίας (98% & 92% αντίστοιχα για κάθε δείγμα). Ωστόσο για το 2ο δείγμα δεδομένων θα περιμέναμε μια μεγαλύτερη αύξηση στο ποσοστό επιτυχίας. Αφού το δεδομένα εισόδου αυξήθηκαν από

75 σε 190 μετρήσεις, θα ήταν αναμενόμενο να δούμε και μια μεγαλύτερη αύξηση στο ποσοστό επιτυχίας απ' ότι τώρα (από 90% σε 92%).

C4.5 (Weka)

Στο δεύτερο πείραμα εξετάζουμε τα αποτελέσματα κατασκευής δένδρου με τη χρήση του αλγορίθμου C4.5 που δίνεται από το εργαλείο Weka.

Πίνακας 10: Αποτελέσματα C4.5 χωρίς χρήση PCA

Πλήθος Δεδομένων Εκπαίδευσης	Ποσοστό Επιτυχίας
1^{ο} Δείγμα	Προβλέψεις Μετρήσεων 251 – 300
20 (1^{ο} Δένδρο)	64%
75 (2^{ο} Δένδρο)	100%
190 (3^{ο} Δένδρο)	92%
2^{ο} Δείγμα	Προβλέψεις Μετρήσεων 338-387
20 (1^{ο} Δένδρο)	42%
75 (2^{ο} Δένδρο)	92%
190 (3^{ο} Δένδρο)	92%

Τα ποσοστά επιτυχίας του αλγορίθμου C4.5 (Weka) είναι αρκετά υψηλά και αρκετά κοντά σε σχέση με τον Decision Tree (Matlab). Ειδικά στο δεύτερο δείγμα δεδομένων υπάρχει ταύτιση των αποτελεσμάτων τόσο στο 1^{ο} όσο και στο 3^{ο} δένδρο. Επίσης συναντάμε και εδώ το φαινόμενο underfiiting για τα δένδρα που κατασκευάζονται με το 1/20 των δεδομένων. Τέλος, παρατηρούμε ότι για τα δένδρα που κατασκευάστηκαν με χρήση του ½ των δεδομένων το ποσοστό επιτυχίας μειώνεται στο πρώτο δείγμα και παραμένει σταθερό στο δεύτερο. Οι αποκλίσεις αυτές, οφείλονται στο λεγόμενο overfitting. Πιο συγκεκριμένα, η συνάρτηση πρόβλεψης λαμβάνει υπόψη τον θόρυβο που εμπεριέχεται σε παλαιότερες μη-αντιπροσωπευτικές τιμές του εκάστοτε χαρακτηριστικού και γι' αυτό αντί να αυξάνεται το ποσοστό επιτυχίας, μειώνεται.

M5P (Weka)

Στο τρίτο πείραμα εξετάζουμε τα αποτελέσματα κατασκευής δένδρου με τη χρήση του αλγορίθμου M5P που δίνεται από το εργαλείο Weka.

Πίνακας 11: Αποτελέσματα M5P χωρίς χρήση PCA

Πλήθος Μετρήσεων για Δεδομένα Εκπαίδευσης	Ποσοστό Επιτυχίας
1^{ο} Δείγμα	Προβλέψεις Μετρήσεων 251 – 300
20 (1^{ο} Δένδρο)	94%
75 (2^{ο} Δένδρο)	96%
190 (3^{ο} Δένδρο)	94%
2^{ο} Δείγμα	Προβλέψεις Μετρήσεων 338-387
20 (1^{ο} Δένδρο)	54%
75 (2^{ο} Δένδρο)	96%
190 (3^{ο} Δένδρο)	90%

Ο αλγόριθμος M5P δίνει τόσο υψηλά ποσοστά επιτυχίας που κανένας άλλος αλγόριθμος από αυτούς που εξετάζουμε δε δίνει. Αυτό οφείλεται στο γεγονός ότι τα φύλλα από τα δένδρα που δημιουργεί προβλέπουν την τιμή της μέτρησης βάσει μιας γραμμικής συνάρτησης των υπολοίπων μεταβλητών. Οι προηγούμενοι δύο αλγόριθμοι έδιναν ως αποτέλεσμα μια διακριτή τιμή για την θερμοκρασία (a, b, c ή d) ενώ για τον M5P υπολογίζουμε μια πραγματική τιμή βάσει της συνάρτησης και έπειτα εφαρμόζουμε τη διακριτοποίηση για να δούμε σε ποια κατηγορία ανήκει κάθε τιμή. Το ποσοστό του 1^{ου} δένδρου για το δεύτερο δείγμα δεδομένων είναι σχετικά χαμηλό για την απόδοση του συγκεκριμένου αλγορίθμου αλλά οφείλεται στο underfitting. Πιο συγκεκριμένα ζητάμε από τον αλγόριθμο να προβλέψει ένα σύνολο μετρήσεων με αρκετά χαμηλές τιμές, ενώ τα δεδομένα εισόδου καλύπτουν μεγαλύτερες τιμές θερμοκρασίας και συνεπώς δεν είναι αντιπροσωπευτικά.

Decision Stump (Weka)

Στο τέταρτο πείραμα εξετάζουμε τα αποτελέσματα κατασκευής δένδρου με τη χρήση του αλγορίθμου Decision Stump που δίνεται από το εργαλείο Weka.

Πίνακας 12: Αποτελέσματα Decision Stump χωρίς χρήση PCA

Πλήθος Δεδομένων Εκπαίδευσης	Ποσοστό Επιτυχίας
1^{ο} Δείγμα	Προβλέψεις Μετρήσεων 251 – 300
20 (1^{ο} Δένδρο)	52%
75 (2^{ο} Δένδρο)	36%
190 (3^{ο} Δένδρο)	34%

2ο Δείγμα	Προβλέψεις Μετρήσεων 338-387
20 (1ο Δένδρο)	60%
75 (2ο Δένδρο)	10%
190 (3ο Δένδρο)	84%

Τα αποτελέσματα του αλγορίθμου Decision Stump είναι αρκετά χαμηλά. Σε καμία περίπτωση δεν μπορούν να συγκριθούν με τα υψηλά ποσοστά που έδωσαν οι προηγούμενοι αλγόριθμοι. Ο λόγος που προκύπτουν αυτά τα αποτελέσματα είναι διότι το δένδρο που κατασκευάζεται είναι ενός επιπέδου και αποτελείται μόνο από την ρίζα και δύο φύλλα (σε όλες τις περιπτώσεις). Μάλιστα τα δύο φύλλα που το αποτελούν δίνουν ως αποτέλεσμα μια σταθερή τιμή πρόβλεψης. Άρα δεν είναι δυνατόν σε καμία περίπτωση να καταφέρει να προβλέψει παραπάνω από δύο κατηγορίες θερμοκρασίας. Το παράδοξο είναι ότι για το 3ο δένδρο του δεύτερου δείγματος τα αποτελέσματα είναι αρκετά ικανοποιητικά και οφείλεται στην ομοιογένεια των δεομένων εισόδου. Γενικά είναι προτιμότερο να χρησιμοποιείται για να προβλέψει λίγες τιμές και λειτουργεί καλύτερα με λίγα και αντιπροσωπευτικά δεδομένα εισόδου. Ενδείκνυται σε περιπτώσεις που το μέγεθος του δένδρου οφείλει να είναι το ελάχιστο δυνατό αλλά δεν υπάρχει μεγάλη αξιοπιστία στα αποτελέσματά του.

RepTree (Weka)

Στο πέμπτο πείραμα εξετάζουμε τα αποτελέσματα κατασκευής δένδρου με τη χρήση του αλγορίθμου RepTree που δίνεται από το εργαλείο Weka.

Πίνακας 13: Αποτελέσματα RepTree χωρίς χρήση PCA

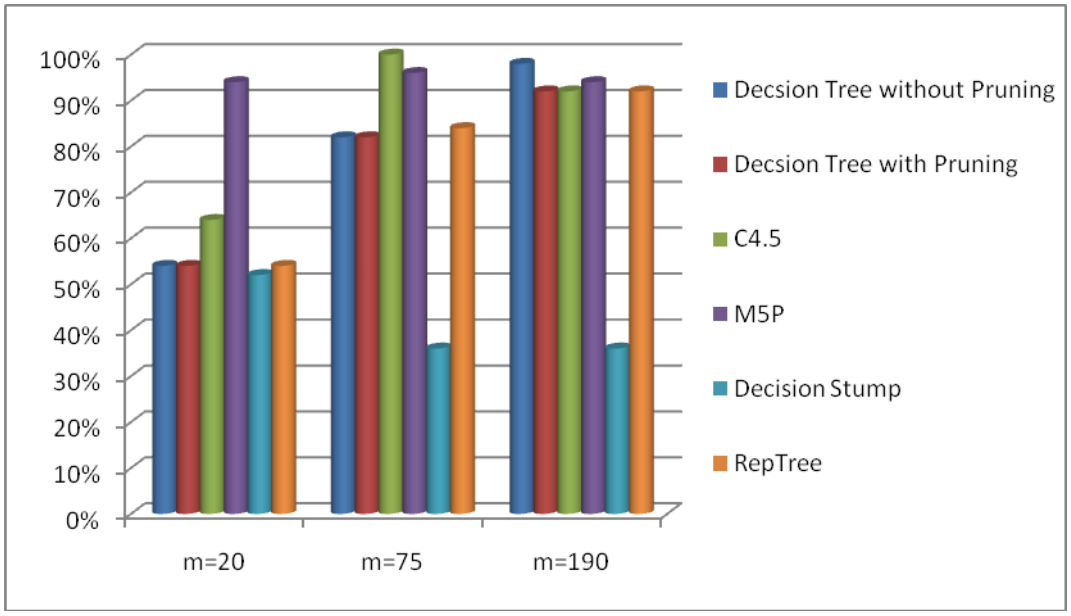
Πλήθος Δεδομένων Εκπαίδευσης	Ποσοστό Επιτυχίας
1ο Δείγμα	Προβλέψεις Μετρήσεων 251 – 300
20 (1ο Δένδρο)	54%
75 (2ο Δένδρο)	84%
190 (3ο Δένδρο)	92%
2ο Δείγμα	Προβλέψεις Μετρήσεων 338-387
20 (1ο Δένδρο)	60%
75 (2ο Δένδρο)	78%
190 (3ο Δένδρο)	94%

Ο αλγόριθμος RepTree δίνει αρκετά ικανοποιητικά αποτελέσματα. Τα φύλλα του δένδρου που κατασκευάζει αποτελούνται από πραγματικές τιμές. Με βάση τις τιμές των υπολοίπων γνωρισμάτων υπολογίζουμε την πραγματική τιμή πρόβλεψης και ύστερα εφαρμόζουμε διακριτοποίηση για την αναγωγή της στις κατηγορίες a, b, c και d που έχουμε ορίσει. Τα ποσοστά επιτυχίας για το πρώτο δένδρο κάθε δείγματος είναι

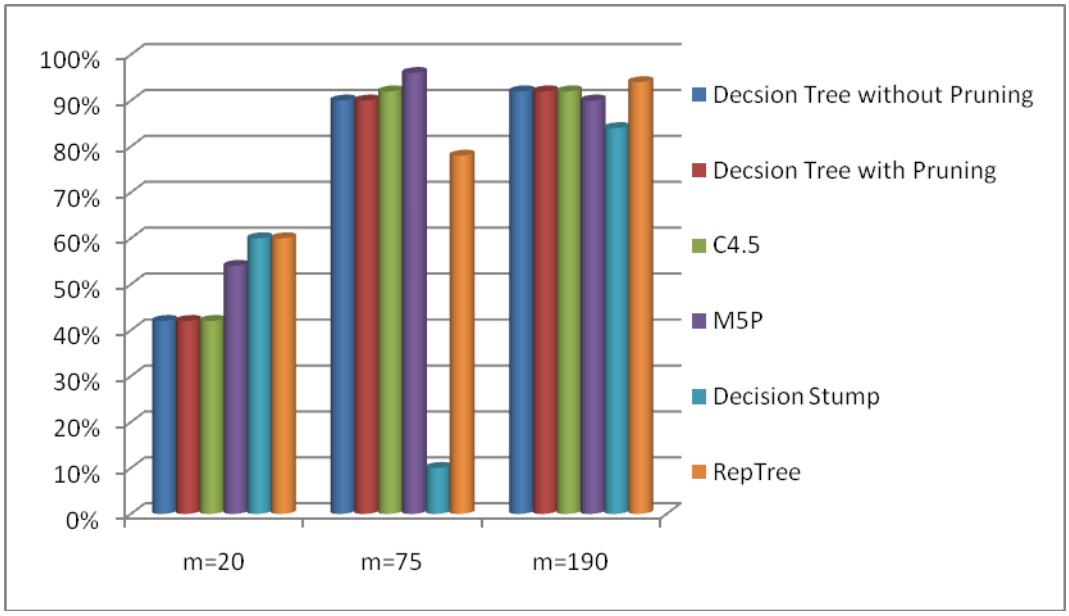
ικανοποιητικά και αντίστοιχα με την πλειονότητα των αλγορίθμων που εξετάσαμε επηρεάζονται και αυτά από το underfitting. Αξιοσημείωτο είναι ότι καθώς το μέγεθος των δεδομένων εκπαίδευσης αυξάνεται, τα ποσοστά επιτυχίας αυξάνονται και αυτά χωρίς να επηρεάζονται από το συσσωρευτικό σφάλμα των προηγούμενων μετρήσεων (overfitting)

Συγκεντρωτική Απόδοση Αλγορίθμων

Παρακάτω ακολουθεί μια διαγραμματική απεικόνιση με τα ποσοστά επιτυχίας όλων των αλγορίθμων. Το πρώτο σχήμα συγκρίνει τα αποτελέσματα των δένδρων που κατασκευάστηκαν με τα δεδομένα που συλλέχθηκαν από τους αισθητήρες για το πρώτο δείγμα δεομένων ενώ στο δεύτερο σχήμα τα ποσοστά επιτυχίας του δεύτερου δείγματος.



Σχήμα 8: Ποσοστά Επιτυχίας Αλγορίθμων χωρίς χρήση PCA για το 1^{ο} δείγμα



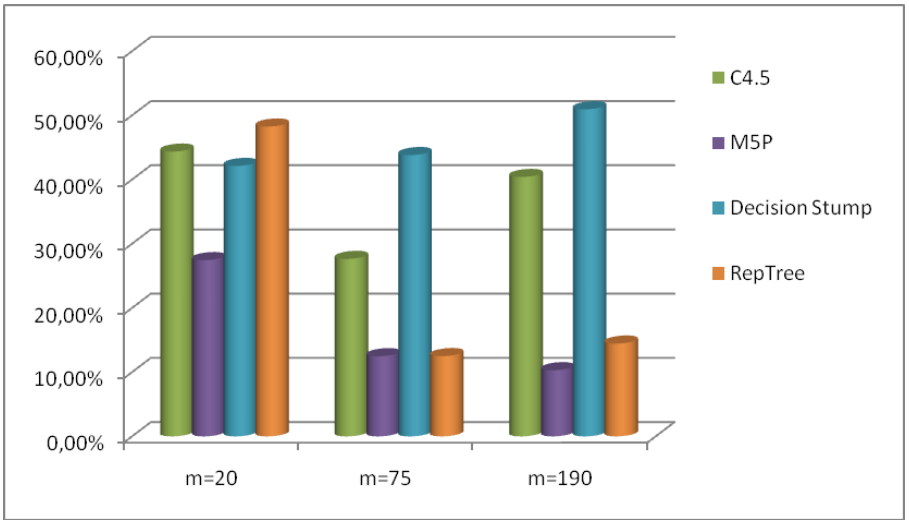
Σχήμα 9: Ποσοστά Επιτυχίας Αλγορίθμων χωρίς χρήση PCA για το 2ο δείγμα

Από τα παραπάνω σχήματα παρατηρούμε ότι ο αλγόριθμος M5P διατηρεί υψηλά επίπεδα απόδοσης και στα δύο δείγματα δεδομένων. Ακολουθεί ο C4.5. Αξιοσημείωτο είναι το γεγονός ότι ο αλγόριθμος Decision Tree, με pruning και χωρίς, έχει τις περισσότερες φορές το ίδιο ποσοστό επιτυχίας. Τα αποτελέσματα του RepTree είναι αξίσου ικανοποιητικά. Ο αλγόριθμος Decision Stump έχει με διαφορά τη χειρότερη απόδοση από όλους τους άλλους και ιδιαίτερα στο δεύτερο δείγμα δεδομένων έχει περίεργη διακύμανση στα ποσοστά επιτυχίας του. Συμπεραίνουμε λοιπόν ότι δεν είναι αξιόπιστος.

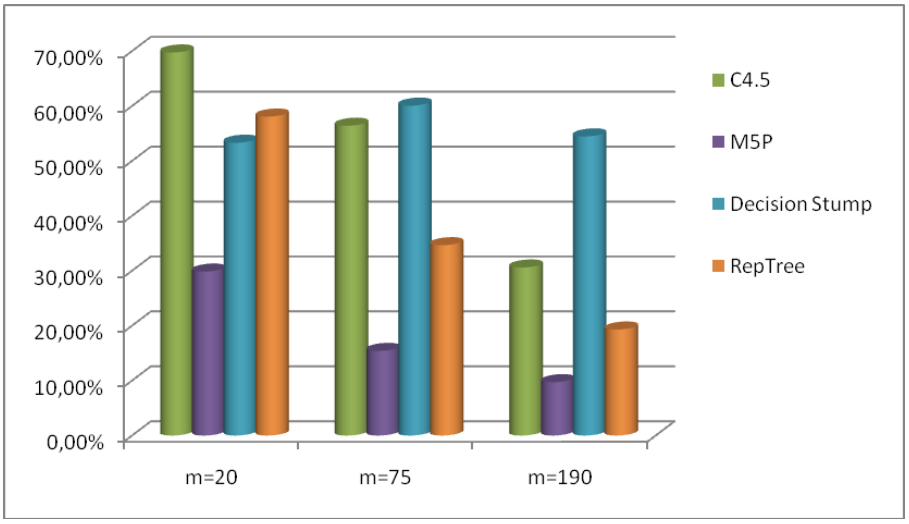
Ακολουθεί ο πίνακας και η διαγραμματική απεικόνιση των ποσοστών της τετραγωνικής ρίζας μέσου τετραγωνικού σφάλματος για τους αλγορίθμους του Weka (C4.5, M5P, Decision Stump & RepTree). Τα ποσοστά βασίζονται στα δεδομένα εκπαίδευσης που χρησιμοποιήθκαν για την κατασκευή του δένδρου απόφασης.

Πίνακας 14: Τετραγωνική Ρίζα Μέσου Τετραγωνικού Σφάλματος (%) χωρίς PCA

Πλήθος Δεδομένων Εκπαίδευσης	Ποσοστό Τετραγωνικής Ρίζας Μέσου Τετραγωνικού Σφάλματος			
	C4.5	M5P	Decision Stump	RepTree
1ο Δείγμα	Προβλέψεις Μετρήσεων 251 – 300			
20 (1ο Δένδρο)	44,36%	27,45%	42,12%	48,26%
75 (2ο Δένδρο)	27,65%	12,49%	43,82%	12,50%
190 (3ο Δένδρο)	40,41%	10,31%	50,95%	14,46%
2ο Δείγμα	Προβλέψεις Μετρήσεων 338-387			
20 (1ο Δένδρο)	69,74%	29,85%	53,32%	58,07%
75 (2ο Δένδρο)	56,41%	15,41%	60,06%	34,65%
190 (3ο Δένδρο)	30,59%	9,71%	54,47%	19,29%



Σχήμα 10: Ρίζα Μέσου Τετραγωνικού Σφάλματος 1ου δείγματος χωρίς PCA



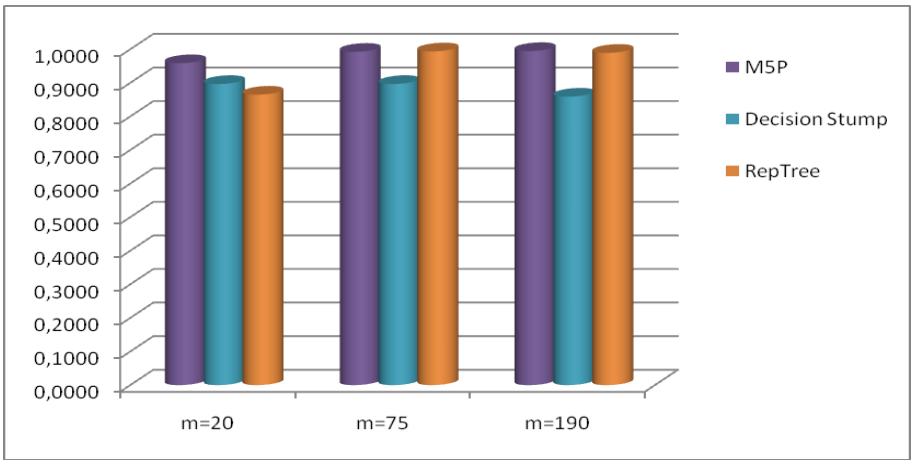
Σχήμα 11: Ρίζα Μέσου Τετραγωνικού Σφάλματος 2ου δείγματος χωρίς PCA

Τα παραπάνω διαγράμματα επιβεβαιώνουν την αξιοπιστία του αλγορίθμου M5P αφού το σφάλμα είναι κατά πολύ μικρότερο από τους υπόλοιπους αλγορίθμους σε όλες τις περιπτώσεις. Όπως έδειξαν και τα ποσοστά επιτυχίας, το μεγαλύτερο σφάλμα κατά την κατασκευή του δένδρου αντιστοιχεί στον Decision Stump. Ο RepTree δείχνει μια καλή συμπεριφορά. Μπορεί το σφάλμα να ξεκινάει από υψηλά ποσοστά όταν το δεδομένα εκπαίδευσης είναι λίγα αλλά μειώνεται ραγδαία καθώς αυξάνεται η είσοδος των δεδομένων. Τέλος, το σφάλμα μειώνεται και για τον C4.5 καθώς αυξάνεται το πλήθος των δεδομένων εισόδου παρόλα αυτά εξακολουθεί να βρίσκεται σε υψηλά επίπεδα.

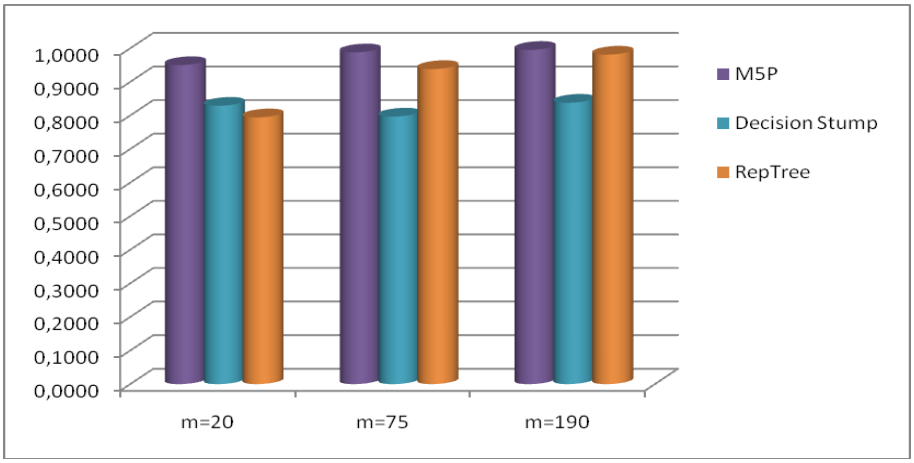
Για τους αλγόριθμους που δίνουν ως έξοδο πραγματική τιμή πρόβλεψης, παρατίθεται πίνακας με τον συντελεστή συσχέτισης που προέκυψε από τα δεδομένα εκπαίδευσης. Οι συντελεστές του M5P πλησιάζουν πάρα πολύ την μονάδα συνεπώς τα φύλλα του δένδρου συσχετίζονται πάρα πολύ με τις πραγματικές τιμές που είχε ως είσοδο. Ακολουθεί ο RepTree με αρκετά υψηλούς συντελεστές ενώ τους χαμηλότερους συντελεστές έχει ο Decision Stump.

Πίνακας 15: Συντελεστές Συσχέτισης Αριθμητικών Αλγορίθμων χωρίς PCA

Πλήθος Δεδομένων Εκπαίδευσης	Συντελεστής Συσχέτισης		
	M5P	Decision Stump	RepTree
1ο Δείγμα	Προβλέψεις Μετρήσεων 251 – 300		
20 (1ο Δένδρο)	0,9586	0,8966	0,8649
75 (2ο Δένδρο)	0,9924	0,8967	0,9937
190 (3ο Δένδρο)	0,9947	0,8589	0,9894
2ο Δείγμα	Προβλέψεις Μετρήσεων 338-387		
20 (1ο Δένδρο)	0,9500	0,8289	0,7943
75 (2ο Δένδρο)	0,9878	0,7968	0,9383
190 (3ο Δένδρο)	0,9953	0,8379	0,9812



Σχήμα 12: Συντελεστές Συσχέτισης 1ου δείγματος χωρίς PCA



Σχήμα 13: Συντελεστές Συσχέτισης 2ου δείγματος χωρίς PCA

5.3.4 Παραδείγματα με χρήση PCA

Η μεθοδολογία PCA εφαρμόστηκε στους αλγορίθμους πρόβλεψης πραγματικών τιμών (M5P, Decision Stump, RepTree). Φυσικά μπορεί να εφαρμοστεί και στους αλγορίθμους πρόβλεψης διακριτών τιμών αλλά ο σκοπός των πειραμάτων είναι να υπολογιστεί η απώλεια πληροφορίας έπειτα από την εφαρμογή τεχνικών μείωσης διαστάσεων. Για το λόγο αυτό υπολογίσαμε τη μέση τιμή του σχετικού σφάλματος ανάμεσα στην αρχική πρόβλεψη και την πρόβλεψη με PCA με χρήση του τύπου :

$\frac{|p_i - p_j|}{|p_i|}$, όπου p_i αρχική πρόβλεψη και p_j πρόβλεψη με pca.

Υστερα υπολογίσαμε την διακύμανση του σχετικού σφάλματος βάσει του τύπου:

$\frac{\sum (x - \bar{x})^2}{n-1}$, όπου $\bar{x}$ είναι η μέση τιμή του σχετικού σφάλματος και

n είναι το μέγεθος του δείγματος.

Η χρήση της μεθοδολογίας PCA δημιούργησε έξι κύριες συνιστώσες που αντικαθιστούν τις έξι μεταβλητές εισόδου (temp_1, hum_1, temp_2, hum_2, hum_3, wind_speed). Για να μειωθεί το μέγεθος των αρχικών δεδομένων χρησιμοποιούμε μόνο τις τρείς πρώτες κύριες συνιστώσες, οι οποίες καλύπτουν πάνω από το 95% της αρχικής πληροφορίας.

Ακολουθούν οι πίνακες με τα αποτελέσματα από κάθε αλγόριθμο ξεχωριστά και για τα δυο δείγματα δεδομένων.

Πίνακας 16: Αποτελέσματα M5P με χρήση PCA

	M5P		
	m=20	m=75	m=190
1ο Δείγμα	Προβλέψεις Μετρήσεων 251 – 300		
Μέση Τιμή Σχετικού Σφάλματος	0,228668	0,054694	0,048628
Διακύμανση Σχετικού Σφάλματος	0,018824	0,002917	0,001912
2ο Δείγμα	Προβλέψεις Μετρήσεων 338-387		
Μέση Τιμή Σχετικού Σφάλματος	0,008488	0,314408	0,188566
Διακύμανση Σχετικού Σφάλματος	0,000016	0,013708	0,121981

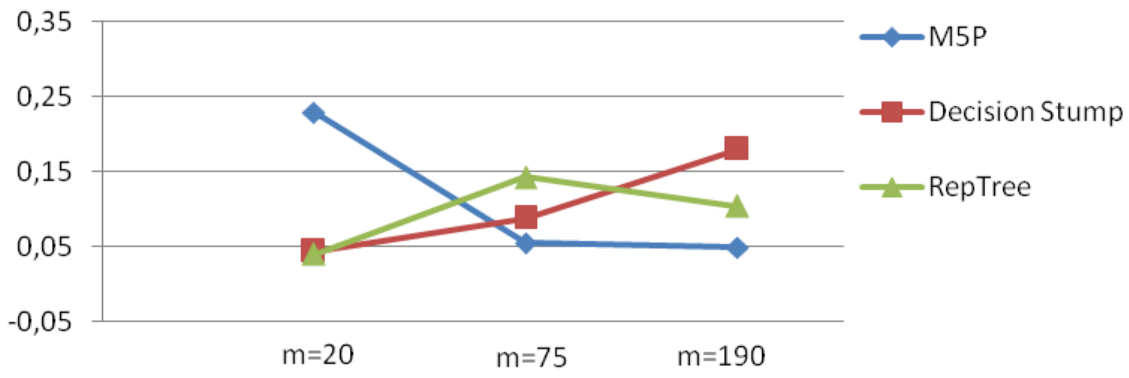
Πίνακας 17: Αποτελέσματα Decision Stump με χρήση PCA

	Decision Stump		
	m=20	m=75	m=190
1ο Δείγμα	Προβλέψεις Μετρήσεων 251 – 300		
Μέση Τιμή Σχετικού Σφάλματος	0,043291	0,087465	0,179494
Διακύμανση Σχετικού Σφάλματος	0,004781	0,04044	0,117258
2ο Δείγμα	Προβλέψεις Μετρήσεων 338-387		
Μέση Τιμή Σχετικού Σφάλματος	0,036253	0,036342	0,312641
Διακύμανση Σχετικού Σφάλματος	0,004009	0,014255	0,102885

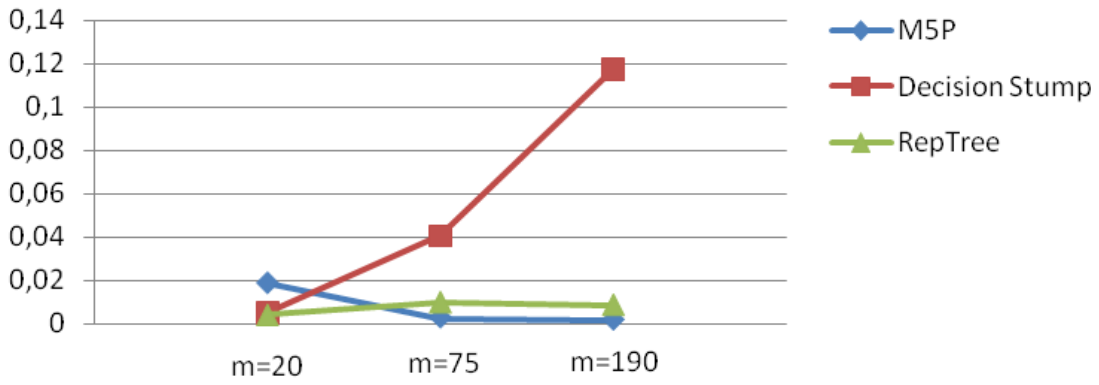
Πίνακας 18: Αποτελέσματα RepTree με χρήση PCA

	RepTree		
	m=20	m=75	m=190
1ο Δείγμα	Προβλέψεις Μετρήσεων 251 – 300		
Μέση Τιμή Σχετικού Σφάλματος	0,039406	0,142243	0,103521
Διακύμανση Σχετικού Σφάλματος	0,004075	0,010025	0,008816
2ο Δείγμα	Προβλέψεις Μετρήσεων 338-387		
Μέση Τιμή Σχετικού Σφάλματος	0,056500	0,219002	0,222628
Διακύμανση Σχετικού Σφάλματος	0,006361	0,019145	0,020612

Στη συνέχεια ακολουθεί μια συγκεντρωτική διαγραμματική απεικόνιση όλων των αλγορίθμων με την μέση τιμή του σφάλματος και τη διακύμανση για κάθε δείγμα.

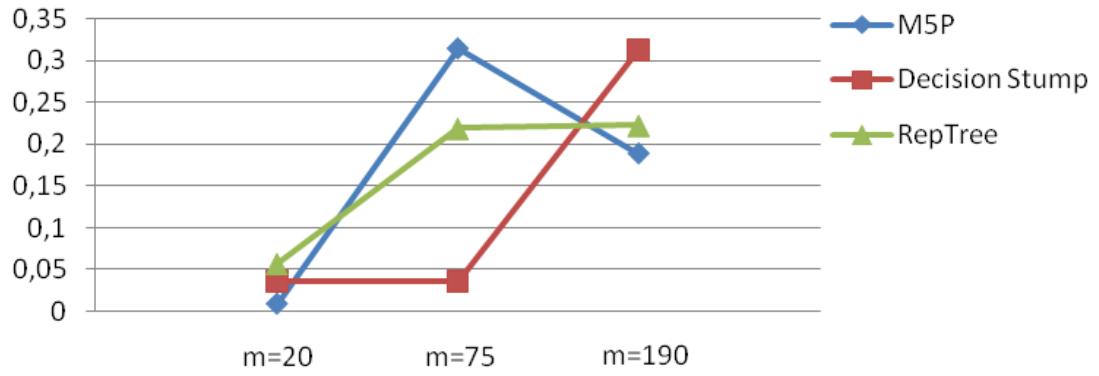


Σχήμα 14: Μέση Τιμή Σχετικού Σφάλματος 1ου δείγματος

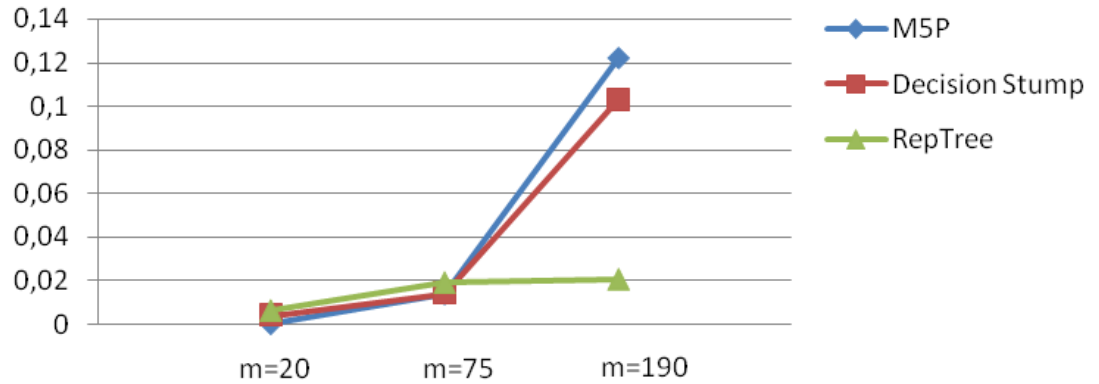


Σχήμα 15: Διακύμανση Σχετικού Σφάλματος 1ου δείγματος

Όπως παρατηρούμε από το 1ο δείγμα, η μέση τιμή του σφάλματος για τον αλγόριθμο M5P ξεκινάει από μια υψηλή τιμή συγκριτικά με τους υπόλοιπους αλγορίθμους όταν το σύνολο δεδομένων εκπαίδευσης είναι μικρό αλλά μειώνεται σημαντικά καθώς τα δεδομένα εκπαίδευσης αυξάνονται. Άρα όσο περισσότερα δεδομένα εκπαίδευσης εισάγονται η απώλεια πληροφορίας με την χρήση PCA μειώνεται σταδιακά. Δεν ισχύει το ίδιο και για τους άλλους δύο αλγορίθμους, όπου η απώλεια πληροφορίας αυξάνεται με την αύξηση των δεδομένων. Ειδικά στον αλγόριθμο Decision Stump αυξάνεται αρκετά το σφάλμα αλλά και η διακύμανση του, υποδεικνύοντας μεγάλες αποκλίσεις στην αρχική τιμή πρόβλεψης και στην πρόβλεψη που προκύπτει από την εφαρμογή PCA.



Σχήμα 16: Μέση Τιμή Σχετικού Σφάλματος 2ου δείγματος



Σχήμα 17: Διακύμανση Σχετικού Σφάλματος 2ου δείγματος

Όσον αφορά το 2^{ο} δείγμα, το σφάλμα κυμαίνεται σε μεγαλύτερα επίπεδα. Ωστόσο, οι αλγόριθμοι Decision Stump και RepTree, διατηρούν την ίδια συμπεριφορά σε σχέση με το 1^{ο} δείγμα. Πιο συγκεκριμένα, για τον αλγόριθμο Decision Stump, όταν m=20 (πλήθος εκπαιδευτικών δεδομένων) το σφάλμα και η διακύμανση είναι αρκετά χαμηλά, για m=75 υπάρχει μια μικρή αύξηση ενώ για m=190 τόσο το σφάλμα όσο και η διακύμανση εκτοξεύονται σε υψηλά επίπεδα. Ο RepTree, διατηρεί πολύ χαμηλή διακύμανση σε όλες τις περιπτώσεις, αν και το σφάλμα αυξάνεται καθώς τα δεδομένα εκπαίδευσης πληθαίνουν. Ο M5P δεν έχει σταθερή πορεία καθώς το σφάλμα και η διακύμανση αυξομειώνονται ανάλογα με το πλήθος των δεδομένων εισόδου.

Είναι αναμενόμενο να αυξάνεται το σφάλμα και η διακύμανση καθώς αυξάνεται το m γιατί τα δεδομένα που χρησιμοποιούνται για την δημιουργία των κύριων συνιστωσών είναι περισσότερα και με μεγαλύτερες αποκλίσεις. Αυτό συνεπάγεται μεγαλύτερη απώλεια πληροφορίας κατά τον μετασχηματισμό. Παρατηρούμε ότι το ίδιο ισχύει και στο 1^{ο} δείγμα αλλά σε μικρότερο βαθμό. Δηλαδή, το σφάλμα και η διακύμανση αυξάνονται καθώς αυξάνεται το m αλλά με πολύ μικρότερο ρυθμό απ'ότι στο δεύτερο δείγμα. Ο ρυθμός αύξησης επηρεάζεται από την αναλογία των τιμών που πρέπει να προβλέψει ο κάθε αλγόριθμος σε σχέση με τις τιμές που έχουν δοθεί ως δεδομένα εκπαίδευσης. Για παράδειγμα, στο 2^{ο} δείγμα όπου m=75, οι αλγόριθμοι M5P και RepTree παρουσιάζουν μια παράδοξη αύξηση. Αυτή οφείλεται στο ότι οι τιμές της θερμοκρασίας που πρέπει να υπολογιστούν είναι σχετικά χαμηλές ενώ το εκπαιδευτικό σύνολο περιελάμβανε μικρό ποσοστό αντίστοιχων μετρήσεων.

Συνολικά, ο αλγόριθμος RepTree φαίνεται να είναι πιο σταθερός συγκριτικά με τους υπόλοιπους αλγόριθμους γιατί διατηρεί χαμηλό σφάλμα και ακόμα χαμηλότερη διακύμανση. Στα τέσσερα από τα έξι πειράματα που έγιναν με χρήση PCA ο M5P είχε καλύτερα αποτελέσματα. Οι αλγόριθμοι M5P και Decision Stump παρουσιάζουν αυξομειώσεις στο σφάλμα και τη διακύμανση αλλά σε γενικές γραμμές οι αποκλίσεις από τις αρχικές προβλέψεις δεν είναι υψηλές. Δηλαδή, όταν το σχετικό σφάλμα είναι περίπου 30% (περιπτώσεις M5P για m=75, Decision Stump για m=190) θεωρείται αρκετά υψηλό αλλά δε θα πρέπει να μας προκαλεί ανησυχία γιατί αυτό το ποσοστό αντιστοιχεί σε μια διαφορά περίπου 3,5 βαθμών °C από την αρχική πρόβλεψη.

6 ΣΥΜΠΕΡΑΣΜΑΤΑ – ΑΝΟΙΧΤΑ ΘΕΜΑΤΑ

6.1 Συμπεράσματα

Στην παρούσα διπλωματική εργασία μελετήθηκαν αλγόριθμοι που μπορούν να χρησιμοποιηθούν για την εκτίμηση ελλιπούς πληροφορίας σε ασύρματα δίκτυα αισθητήρων. Ένας σταθμός βάσης που συλλέγει πληροφορίες από αισθητήρες, μπορεί να χρησιμοποιήσει αυτές τις μεθοδολογίες προκειμένου να αποφύγει τις επαναλήψεις εκπομπών των απολεσθέντων τιμών. Με την αντικατάσταση της ελλιπούς τιμής από την τιμή πρόβλεψης εξοικονομούμε ενέργεια αποφεύγοντας την υπολογιστική συμφόρηση του σταθμού βάσης από τις επαναλαμβανόμενες εκπομπές και απαλλάσσοντας τον από την διαδικασία διαχείρισης διπλών τιμών.

Το δείγμα που χρησιμοποιήθηκε στις στατιστικές μεθοδολογίες, παρουσιάζει υψηλό δείκτη συσχέτισης. Συνεπώς τα συμπεράσματα δεν μπορούν να γενικευτούν σε δείγμα με χαμηλή συσχέτιση μεταβλητών. Επίσης το μέγεθος του δείγματος είναι πιθανό να τροποποιήσει τα αποτελέσματα των πειραμάτων. Σε ένα πολύ μεγαλύτερο δείγμα ενδέχεται να προκύψουν διαφορετικά αποτελέσματα καθώς η συσχέτιση των μεταβλητών είναι πιθανό να διαφοροποιείται.

Τα πειράματα που έγιναν επικεντρώθηκαν στην πρόβλεψη τιμών μέσα από τον τομέα της στατιστικής ανάλυσης με χρήση της μεθοδολογίας extrapolation και μέσα από αλγορίθμους κατηγοριοποίησης.

Η μεθοδολογία extrapolation έδειξε ότι χρειάστηκαν περίπου 10 μετρήσεις για να σταθεροποιηθεί η τιμή μιας πρόβλεψης. Όσο μικρότερο ήταν το δείγμα εκπαίδευσης τόσο καλύτερες ήταν οι τιμές εκτίμησης. Αυτό οφείλεται στο γεγονός ότι σε ένα σχετικά μικρό δείγμα οι μεταβολές ανάμεσα στις τιμές των μεταβλητών είναι μικρότερες, οπότε υπάρχει μεγαλύτερη συσχέτιση, και συνεπώς η ακρίβεια των αποτελεσμάτων είναι καλύτερη. Σε σχέση με τις προβλέψεις που έγιναν με βάση την μέση τιμή παρατηρήθηκε ότι η μεθοδολογία extrapolation μειονεκτεί.

Όσον αφορά τους αλγορίθμους κατηγοριοποίησης, τα δένδρα που κατασκευάστηκαν με χρήση pruning δεν διέφεραν στις προβλέψεις τους απ’ότι τα δένδρα χωρίς pruning. Στα πειράματα που χρησιμοποιήθηκε μεγαλύτερο ποσοστό δεδομένων ως δεδομένα εκπαίδευσης τα αποτελέσματα των αλγορίθμων είχαν καλύτερη απόδοση τις περισσότερες φορές. Ωστόσο, σε αρκετές περιπτώσεις η αναλογία αύξησης του ποσοστού των δεδομένων εκπαίδευσης ως προς τη βελτίωση των αποτελεσμάτων ήταν δυσανάλογη. Δηλαδή, παρότι το ποσοστό των δεδομένων εκπαίδευσης αυξήθηκε αρκετά, το ποσοστό επιτυχίας δεν αυξήθηκε αναλόγως. Πρέπει λοιπόν να αναρωτηθούμε εάν αξίζει να υποστούμε το κόστος της αύξησης των δεδομένων εκπαίδευσης σε σχέση με τι μπορεί να προσφέρει αυτό. Η χρήση της τεχνικής μείωσης των αρχικών διαστάσεων PCA οδήγησε σε μικρή απώλεια πληροφορίας σε σχέση με τις αρχικές προβλέψεις. Στα περισσότερα πειράματα οι αποκλίσεις ήταν αρκετά μικρές. Ιδιαίτερα όταν τα δεδομένα εκπαίδευσης ήταν λίγα το σφάλμα ανάμεσα στην αρχική πρόβλεψη και την πρόβλεψη με χρήση PCA ήταν σχεδόν αμελητέο.

Ανοιχτά Θέματα

Μέσα από τα πειράματα που υλοποιήθηκαν, έγινε μια προσπάθεια να μελετηθούν όλες οι παράμετροι που επηρεάζουν την απόδοση των αλγορίθμων. Ωστόσο παραμένουν ακόμα και άλλοι παράγοντες που δεν έχουν μελετηθεί επαρκώς και θα μπορούσαν να είναι αντικείμενα επιπρόσθετης μελέτης.

Ορισμένα ανοιχτά θέματα που αξίζει να μελετηθούν περαιτέρω είναι τα εξής:

- Η χρήση διαφορετικών τεχνικών μείωσης διαστάσεων μπορεί να μεταβάλει σε μεγάλο βαθμό τα αποτελέσματα των πειραμάτων. Στο 2^{ο} κεφάλαιο αναφέρονται οι τεχνικές PCA και Factor Analysis, ωστόσο η δεύτερη τεχνική δεν χρησιμοποιήθηκε στα πειράματα των αλγορίθμων αν και αξίζει να μελετηθούν τα αποτελέσματα που θα προέκυπταν και η σύγκριση τους με αυτά της PCA.

- Η αλλαγή στον τρόπο υπολογισμού των κύριων συνιστωσών καθώς και στον καθορισμό του αριθμού των κύριων συνιστωσών μπορεί να βελτιώσει τον χρόνο υπολογισμού της εκτίμησης της ελλιπούς τιμής.

- Η μελέτη όλων των κατηγοριών σφαλμάτων που πιθανώς να προκύψουν σε ένα ασύρματο δίκτυο αισθητήρων. Στο 1^{ο} κεφάλαιο αναφέρονται αναλυτικά όλες οι περιπτώσεις. Στα πλαίσια της μελέτης, ερευνήθηκαν οι περιπτώσεις μετάδοσης μεμονωμένων τιμών καθώς και η πλήρη αποτυχία μετάδοσης τιμών. Θα είχε ενδιαφέρον να μελετηθεί η απόδοση των αλγορίθμων και των τεχνικών που χρησιμοποιήθηκαν σε περιπτώσεις υποβάθμισης ακρίβειας των μετρήσεων, σε συσσωρευτική απόκλιση σφαλμάτων στην τιμή που εκπέμπεται καθώς και σε δεδομένα που αποκλίνουν κατά μια σταθερά από την πραγματική τιμή.

- Ο τρόπος με τον οποίο θα ανανεώνεται δυναμικά κάθε μοντέλο πρόβλεψης, προκειμένου να είναι συνεχώς ενημερωμένο. Κατά τη διάρκεια των πειραμάτων, αρχικά δημιουργούσαμε ένα μοντέλο βάσει των δεδομένων εκπαίδευσης και ύστερα μετρούσαμε το ποσοστό επιτυχίας κάθε αλγορίθμου. Σε πραγματικό χρόνο όμως, η ροή των δεδομένων είναι συνεχής οπότε το αρχικό μοντέλο πρέπει να ανανεώνεται ανά τακτά χρονικά διαστήματα ώστε να ανταποκρίνεται στα πιο πρόσφατα δεδομένα.

ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

Ξενόγλωσσος όρος	**Ελληνικός Όρος**
Pervasive Computing	Διάχυτος Υπολογισμός
Ubiquitous Computing	Διάχυτος Υπολογισμός
Hardware	Υλικό
Software	Λογισμικό
Mainframe Era	Περίοδος Μεγάλων Συστημάτων Υπολογιστών
Personal Computer Era	Περίοδος Προσωπικού Υπολογιστή
Ubiquitous Computing Era	Περίοδος Διάχυτης Πληροφόρησης
Distributed Systems	Κατανεμημένα Συστήματα
Mobile Computing	Κινητός Υπολογισμός
Smart Space	Ευφυής Χώρος
Invisibility	Αορατότητα
Localized Scalability	Δυνατότητα Κλιμάκωσης
Masking Uneven Conditioning	Απόκρυψη Διαφορετικών Συνθηκών Περιβάλλοντος
Context-Aware Computing	Πληροφορία Πλαισίου
Interpolation	Παρεμβολής
Extrapolation	Προεκβολής
Feature Selection	Επιλογή Γνωρισμάτων
Feature Extraction	Εξαγωγή Γνωρισμάτων
Common Factor Analysis	Παραγοντική Ανάλυση
Principal Component Analysis	Ανάλυση Κύριων Συνιστωσών
Principal Components	Κύριες Συνιστώσες
Covariance Matrix	Πίνακας Συνδιακύμανσης
Smart Sensor	Έξυπνος Αισθητήρας
Global Position System	Συστήματα Εντοπισμού Θέσης
Environmental Observations and Forecasting Systems	Συστήματα Παρατήρησης Περιβάλλοντος και Πρόβλεψης
Wireless Sensor Network	Ασύρματο Δίκτυο Αισθητήρων
Base Station	Σταθμό Βάσης
Regression	Παλινδρόμηση
Partial Least Square Regression	Μερική Παλινδρόμηση Ελαχίστων Τετραγώνων
Noise	Θόρυβος
Classification	Κατηγοριοποίηση
Training Data	Δεδομένα Εκμάθησης
Decision Tree	Δένδρο Απόφασης
Information Gain	Κέρδος Πληροφορίας
Pruning	Κλάδεμα
Regression Tree	Δένδρο Παλινδρόμησης
Model Tree	Πρότυπο Δένδρο
Threshold	Όριο
Clustering	Ομαδοποίηση

Association Rule Mining	Εξαγωγή Κανόνων Συσχέτισης
Attribute Selection	Επιλογή Γνωρισμάτων
Nearest Neighbour Interpolation	Παρεμβολή πλησιέστερου γείτονα
Linear Interpolation	Γραμμική Παρεμβολή
Grid Points	Σημείο Πλέγματος
Root Mean Squared Error	Τετραγωνική Ρίζα Μέσου Τετραγωνικού Σφάλματος
Correlation Coefficient	Συντελεστής Συσχέτισης

ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ

WSN	Wireless Sensor Network
PCA	Principal Component Analysis
ITIS	Intelligent Traffic Information Service
GPS	Global Position System
EOFS	Environmental Observations and Forecasting Systems
CORIE	Columbia River Estuary
SVD	Singular Value Decomposition
E-M	Expectation - Maximization
SECOAS	Self-organizing Collegiate Sensor Networks
NIPALS	Non-linear Iterative Partial Least Squares
PLS	Partial Least Square
WEKA	Waikato Environment for Knowledge Analysis
ARFF	Attribute Relation File Format
JAR	Java Archive
Nan	Not a number

ΑΝΑΦΟΡΕΣ

[1] M. Weiser and J. S. Brown, "The Coming Age of Calm Technology," 1996; http://www.ubiq.com/hypertext/weiser/acmfuture2endnote.htm [Προσπελάστηκε: 30/5/2011]

[2] M. Weiser, "The Computer for the 21st Century", Scientific American, September 1991.

[3] C. Siva Ram Murthy and B. S. Manoj, *Ad Hoc Wireless Networks, Architectures and Protocols*, Prentice-Hall, 2004

[4] CHRONIC Project; http://chronic.cestel.es/ [Προσπελάστηκε: 28/09/2011]

[5] X. Li, J. Wu, X. Lin, Y. Li and M. Li, "ITIS: Intelligent Traffic Information Service in ShanghaiGrid", ChinaGrid Annual Conference, 2008.

[6] F. Akyildiz, W. Su, Y. Sankarasubramaniam and E. Cayirci, "A Survey on Sensor Networks", IEEE Communications Magazine, vol. 40, no. 8, Aug. 2002, pp. 102–14.

[7] E. Della Valle, S. Ceri, F. Harmelen and D. Fensel, "It's a Streaming World! Reasoning upon Rapidly Changing Information", IEEE Intelligent Systems , 24(6):83–89, 2009

[8] J. Hong, E. Suh, J. Kim and S. Kim, "Context-Aware System for Proactive Personalized Service Based on Context History," Expert Systems with Applications, vol. 36, no. 4, 2009, pp. 7448–7457.

[9] R. Dunia, S.J. Qin, T.F. Edgar and T.J. McAvoy, "Identification of Faulty Sensors Using Principal Component Analysis", AIChE Journal, October, vol. 42, no 10, 1996, pp. 2797-2812.

[10] J. Himberg, J. Mäntyjärvi and P. Korpipää, "Using PCA and ICA for Exploratory Data Analysis in Situation Awareness", Proc. IEEE Conf. Multisensor Fusion and Integration for Intelligent Systems, IEEE CS Press, 2001, pp. 127-131.

[11] Mathworks, Matlab Statistic Toolbox, Principal Component Analysis (PCA); http://www.mathworks.com/help/toolbox/stats/brkgqnt.html#f75476 [Προσπελάστηκε: 20/07/2011]

[12] J. E. Jackson, *A User's Guide to Principal Components*, J. Wiley & sons, New York, 1991

[13] D. A. Jackson: "Stopping Rules in Principal Component Analysis: A Comparison of Heuristical and Statistical Approaches", Ecology 74: 2204-2214, 1993.

[14] H. Risvik, "Principal Component Analysis (PCA) & NIPALS algorithm", 2007; http://folk.uio.no/henninri/pca_module/pca_nipals.pdf [Προσπελάστηκε: 20/07/2011]

[15] E. Adams, B. Walczak, C. Vervaet, P.G. Risha and D.L. Massart, "Principal component analysis of dissolution data with missing elements", International Journal of Pharmaceutics, 234, pp.169–178, 2002

[16] R. E. Madsen, L. K. Hansen and O. Winther, "Singular Value Decomposition and Principal Component Analysis", ISP Technical Report, 2004

[17] P. J. Olver, "Orthogonal Bases and the QR Algorithm", Lecture Notes, 2010; http://www.math.umn.edu/~olver/aims_/qr.pdf [Προσπέλαση: 01/08/2011]

[18] I. K. Fodor, "A survey of dimension reduction techniques", Technical Report UCRL-ID-148494, Lawrence Livermore National Laboratory, 2002

[19] J. Himberg, J. Mäntyjärvi and P. Korpipää, "Using PCA and ICA for Exploratory Data Analysis in Situation Awareness", Proc. IEEE Conf. on Multisensor Fusion and Integration for Intelligent Systems, IEEE CS Press, 2001, pp. 127-131.

[20] R. Agarwal and A.R. Rao, "Data Reduction Techniques"; http://www.iasri.res.in/ebook/EB_SMAR/e-book_pdf%20files/Manual%20II/9-data_reduction.pdf [Προσπέλαση: 01/08/2011]

[21] R. L. Burden and D. J. Faires, *Numerical Analysis*, 8th Edition, Thomsom Brooks / Cole, Australia, 2005.

[22] W. Iba and P. Langley, "Induction of One-Level Decision Trees", Proceedings of the 9th International Conference on Machine Learning, pp. 233-240, Morgan Kaufmann Publishers, 1992

[23] J. R. Quinlan, "Learning with Continuous Classes", Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, 1992, pp. 343–348

[24] J. J. Dolado, D. Rodríguez, J. Riquelme, F. Ferrer-Troyano and J. J. Cuadrado, “A Two Stage Zone Regression Method for Global Characterization of a Project Database”, Advances in Machine Learning Application in Software Engineering, Idea Group Publishing, 2007

[25] Y. Wang and I. H. Witten, “Induction of Model Trees for Predicting Continuous Classes”, Proceedings of the Poster Papers of the European Conference on Machine Learning, Prague, 1997, University of Economics, Faculty of Informatics and Statistics

[26] I. H. Witten and E. Frank: “Data Mining, *Practical Machine Learning Tools and Techniques*, Second Edition”, Morgan Kaufmann Publishers, 2005.

[27] J. R. Quinlan: “Induction of Decision Trees”, Machine Learning, v.1 n.1, p.81-106, 1986.

[28] Mathworks, Matlab Statistic Toolbox, Decision Tree; http://www.mathworks.com/products/statistics/demos.html?file=/products/demos/shipping/stats/classdemo.html - 19 [Προσπελάστηκε: 15/09/2011]

[29] V. Tsetsos, N. Silvestros and S. Hadjiefthymiades, "Collaborative Sensing over Smart Sensors", 2nd Student Workshop on Wireless Sensor Netwotks, Athens, 2009.

[30] S. Roweis, “EM Algorithms for PCA and SPCA”, Proceedings of the 1997 conference on Advances in neural information processing systems 10, p.626-632, United States, July 1998

[31] FLOODNET Project; http://envisense.org/floodnet.htm [Προσπελάστηκε: 28/09/2011]

[32] SECOAS Project; http://envisense.org/secoas.htm [Προσπελάστηκε: 28/09/2011]

[33] D.C. Steere, A. Baptista, D. McNamee, C. Pu and J. Walpole, “Research Challenges in Environmental Observations and Forecasting Systems,” Proc. ACM/IEEE Int. Conf. Mobile Computing and Networking (MOBICOMM), 2000, pp. 292-299.

[34] Mathworks, Matlab Interpolation Functions, Technical Document; http://www.mathworks.com/help/techdoc/ref/interp2.html [Προσπελάστηκε: 10/10/2011]