

Πανεπιστήμιο Πειραιώς Τμήμα Πληροφορικής



Αποθήκες Δεδομένων και Εξόρυξη Γνώσης (Data Warehousing & Data Mining)

Προπαρασκευή Δεδομένων

Γιάννης Θεοδωρίδης, Νίκος Πελέκης

Ομάδα Διαχείρισης Δεδομένων
Εργαστήριο Πληροφοριακών Συστημάτων

<http://isl.cs.unipi.gr/db>

Περιεχόμενα



- Προπαρασκευή δεδομένων... Γιατί?
- Εργασίες προπαρασκευής
 - Καθαρισμός δεδομένων
 - Ολοκλήρωση & μετασχηματισμός δεδομένων
 - Μείωση δεδομένων
 - Διακριτοποίηση & παραγωγή ιεραρχιών εννοιών
- Σύνοψη

Προπαρασκευή δεδομένων... Γιατί?



- Τα δεδομένα στον πραγματικό κόσμο είναι «βρώμικα»
 - **ελλιπή**: έλλειψη τιμών χαρακτηριστικών, έλλειψη ενδιαφερόντων χαρακτηριστικών, μόνο αθροιστικά δεδομένα
 - **λανθασμένα**: που περιέχουν λάθη ή outliers
 - **ασυνεπή**: που περιέχουν ασυμφωνίες σε κωδικούς και ονόματα
- Κακή ποιότητα στα δεδομένα → κακή ποιότητα στην εξόρυξη γνώσης!
 - Ποιοτικές αποφάσεις πρέπει να βασίζονται σε ποιοτικά δεδομένα
 - Οι ΑΔ χρειάζονται συνεπή ολοκλήρωση ποιοτικών δεδομένων

Εργασίες Προπαρασκευής Δεδομένων



- Καθαρισμός δεδομένων
 - Συμπλήρωση ελλειπών τιμών, εντοπισμός 'θορύβου', επίλυση ασυνεπειών
- Μετασχηματισμός δεδομένων
 - Εξομάλυνση, κανονικοποίηση τιμών
- Μείωση δεδομένων
 - Μείωση διαστάσεων (χαρακτηριστικών), μείωση πλήθους δεδομένων
- Διακριτοποίηση δεδομένων
 - Από συνεχείς σε διακριτές τιμές ή διαστήματα τιμών

Εργασίες Προπαρασκευής Δεδομένων



- Καθαρισμός δεδομένων
 - Συμπλήρωση ελλειπών τιμών, εντοπισμός 'θορύβου', επίλυση ασυνεπειών
- Μετασχηματισμός δεδομένων
 - Εξομάλυνση, κανονικοποίηση τιμών
- Μείωση δεδομένων
 - Μείωση διαστάσεων (χαρακτηριστικών), μείωση πλήθους δεδομένων
- Διακριτοποίηση δεδομένων
 - Από συνεχείς σε διακριτές τιμές ή διαστήματα τιμών

Ελλιπή δεδομένα



- Τα δεδομένα μπορεί να μην είναι πάντοτε διαθέσιμα
 - Π.χ., πολλές εγγραφές δεν έχουν τιμή για αρκετά χαρακτηριστικά, όπως το εισόδημα ενός πελάτη
- Ελλιπή δεδομένα μπορεί να υπάρχουν για διάφορους λόγους:
 - Μηχανική βλάβη
 - Διαγραφή λόγω ασυνέπειας με άλλα καταγεγραμμένα δεδομένα
 - Καθόλου εισαγωγή λόγω διαφωνίας ή επειδή θεωρήθηκαν ασήμαντα
- Συχνά χρειάζεται να συμπεράνουμε τα ελλιπή δεδομένα

Πώς διαχειριζόμαστε ελλιπή δεδομένα



- Αγνοούμε την εγγραφή
 - συνήθως όταν λείπει η κλάση (αν υποθέσουμε ότι κάνουμε κατηγοριοποίηση - αναποτελεσματική όταν το ποσοστό των τιμών που λείπουν ανά χαρακτηριστικό ποικίλει αρκετά)
- Συμπληρώνουμε την τιμή χειρωνακτικά
 - Κουραστικό – μη πρακτικό!
- Χρήση μίας γενικής σταθεράς για να συμπληρώσουμε την τιμή
 - π.χ., “άγνωστο”
- Χρήση της μέσης τιμής του χαρακτηριστικού
 - Χρήση της μέσης τιμής του χαρακτηριστικού για όλα τα δείγματα που ανήκουν στην ίδια κλάση: εξυπνότερο!
- Χρήση της πιο «πιθανής» τιμής: εξαγωγή συμπεράσματος μέσω Bayesian formula ή δένδρων απόφασης

7

ΠΑ.ΠΕΛ – Γιάννης Θεοδορίδης

Λανθασμένα δεδομένα (‘θόρυβος’)



- Θόρυβος: τυχαίο λάθος ή απόκλιση από μία μετρημένη μεταβλητή
- Λανθασμένες τιμές χαρακτηριστικών μπορεί να οφείλονται σε
 - Ελαττωματικά όργανα συλλογής δεδομένων
 - Προβλήματα κατά την εισαγωγή
 - Προβλήματα κατά τη μετάδοση των δεδομένων
 - Τεχνολογικοί περιορισμοί
 - Ασυνέπεια στην συμφωνημένη ονοματολογία
- Άλλα προβλήματα δεδομένων που απαιτούν καθαρισμό
 - Διπλές εγγραφές
 - Ατελή δεδομένα
 - Ασυνεπή δεδομένα

8

ΠΑ.ΠΕΛ – Γιάννης Θεοδορίδης

Πώς διαχειριζόμαστε λανθασμένα δεδομένα



- Μέθοδος τοποθέτησης σε κουτιά – bins (Binning method):
 - Αρχικά ταξινόμηση δεδομένων και διαμερισμός σε (ίσου μεγέθους) κουτιά
 - κατόπιν εξομάλυνση με τον «μέσο» ή τον «διάμεσο» του κουτιού, εξομάλυνση με τα «όρια» του κουτιού, κ.α.
- Συσταδοποίηση (clustering)
 - Ανακάλυψη και αφαίρεση των «μη φυσιολογικών» δεδομένων (outliers)
- Συνδυασμένη επιθεώρηση από υπολογιστή και άνθρωπο
 - Ανακάλυψη ύποπτων τιμών και έλεγχος από άνθρωπο
- Παλινδρόμηση (regression)
 - Εξομάλυνση μέσω προσαρμογής των δεδομένων σε συναρτήσεις παλινδρόμησης

Binning method



- Διαμερισμός ίσου πλάτους (distance):
 - Χωρίζει την απόσταση σε N διαστήματα ίσου μεγέθους: ομοιόμορφο πλέγμα (uniform grid)
 - Αν A και B είναι οι χαμηλότερες και υψηλότερες τιμές του χαρακτηριστικού, το πλάτος των διαστημάτων θα είναι: $W = (B-A)/N$.
 - Άμεση μέθοδος
 - Τα outliers μπορεί να επιδράσουν στη μέθοδο
 - Αναποτελεσματική διαχείριση ασύμμετρων δεδομένων
- Διαμερισμός ίσου βάθους (frequency):
 - Χωρίζει την απόσταση σε N διαστήματα, καθένα από τα οποία περιλαμβάνει τον ίδιο αριθμό δειγμάτων
 - Καλή κλιμάκωση δεδομένων
 - Δύσκολη διαχείριση κατηγορικών χαρακτηριστικών

Παράδειγμα binning



- * Ταξινομημένες τιμές: 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- * Διαμελισμός σε (ίσου βάθους) bins:
 - Bin 1: 4, 8, 9, 15
 - Bin 2: 21, 21, 24, 25
 - Bin 3: 26, 28, 29, 34
- * εξομάλυνση με το «μέσο» του κουτιού (smoothing by bin means):
 - Bin 1: 9, 9, 9, 9
 - Bin 2: 23, 23, 23, 23
 - Bin 3: 29, 29, 29, 29
- * εξομάλυνση με τα «όρια» του κουτιού (smoothing by bin boundaries):
 - Bin 1: 4, 4, 4, 15
 - Bin 2: 21, 21, 25, 25
 - Bin 3: 26, 26, 26, 34

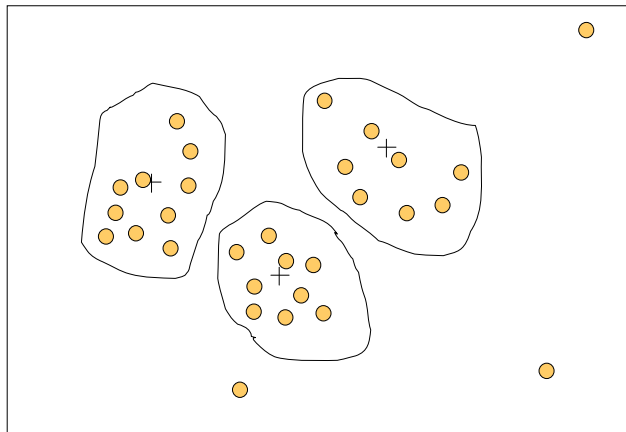
11

ΠΑ.ΠΕΛ – Γιάννης Θεοδορίδης

Ανάλυση συστάδων και εντοπισμός outliers



- Με συσταδοποίηση (με χρήση ειδικών τεχνικών), ομαδοποιούμε τα δεδομένα ανακαλύπτοντας ταυτόχρονα και outliers



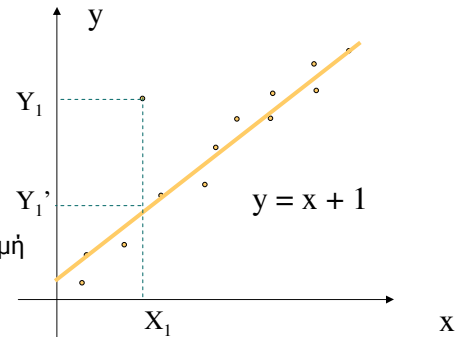
12

ΠΑ.ΠΕΛ – Γιάννης Θεοδορίδης

Παλινδρόμηση



- Ψάχνουμε την πιο αντιπροσωπευτική συνάρτηση για τα δεδομένα μας
- Γραμμική Παλινδρόμηση
 $Y = \alpha + \beta X$
 - Τα δεδομένα μοντελοποιούνται ώστε να ανήκουν σε ευθεία γραμμή
 - Αναζητούμε δύο παραμέτρους, α και β
 - Π.χ. με τη μέθοδο των ελάχιστων τετραγώνων (least squares)
- Τα outliers είναι αυτά που απέχουν «πολύ» από την τιμή που προκύπτει από τη συνάρτηση παλινδρόμησης



13

ΠΑ.ΠΕΛ – Γιάννης Θεοδορίδης

Εργασίες Προπαρασκευής Δεδομένων



- Καθαρισμός δεδομένων
 - Συμπλήρωση ελλειπών τιμών, εντοπισμός 'θορύβου', επίλυση ασυνεπειών
- Μετασχηματισμός δεδομένων
 - Εξομάλυνση, κανονικοποίηση τιμών
- Μείωση δεδομένων
 - Μείωση διαστάσεων (χαρακτηριστικών), μείωση πλήθους δεδομένων
- Διακριτοποίηση δεδομένων
 - Από συνεχείς σε διακριτές τιμές ή διαστήματα τιμών

14

ΠΑ.ΠΕΛ – Γιάννης Θεοδορίδης

Μετασχηματισμός δεδομένων



- Εξομάλυνση: αφαίρεση θορύβου από δεδομένα
- Κανονικοποίηση: κλιμάκωση σε ένα μικρότερο, καθορισμένο εύρος τιμών
 - min-max κανονικοποίηση
 - z-score κανονικοποίηση
 - κανονικοποίηση με δεκαδική κλιμάκωση
- Κατασκευή χαρακτηριστικών
 - Νέα χαρακτηριστικά δημιουργούνται από τα ήδη υπάρχοντα

15

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

Κανονικοποίηση δεδομένων



- min-max κανονικοποίηση

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

- z-score κανονικοποίηση

$$v' = \frac{v - \text{mean}_A}{\text{std Dev}_A}$$

- κανονικοποίηση με δεκαδική κλιμάκωση στο (-1 .. +1)

$$v' = \frac{v}{10^j} \quad \text{όπου } j \text{ ο μικρότερος ακέραιος για τον οποίο } \max(|v'|) < 1$$

16

ΠΑ.ΠΕΛ - Γιάννης Θεοδορίδης

Εργασίες Προπαρασκευής Δεδομένων



- Καθαρισμός δεδομένων
 - Συμπλήρωση ελλειπών τιμών, εντοπισμός 'θορύβου', επίλυση ασυνεπειών
- Μετασχηματισμός δεδομένων
 - Εξομάλυνση, κανονικοποίηση τιμών
- Μείωση δεδομένων
 - Μείωση διαστάσεων (χαρακτηριστικών), μείωση πλήθους δεδομένων
- Διακριτοποίηση δεδομένων
 - Από συνεχείς σε διακριτές τιμές ή διαστήματα τιμών

Στρατηγικές Μείωσης δεδομένων



- Οι ΑΔ αποθηκεύουν terabytes δεδομένων: Πολύπλοκη ανάλυση δεδομένων μπορεί να διαρκέσει πολύ χρόνο για όλο τον όγκο των δεδομένων
- Μείωση δεδομένων
 - Με είσοδο μία μειωμένη αναπαράσταση του συνόλου δεδομένων πολύ μικρότερη σε όγκο παράγονται τα ίδια (ή σχεδόν τα ίδια) αναλυτικά αποτελέσματα
- Στρατηγικές μείωσης
 - Συσσώρευση κύβων
 - Μείωση διαστάσεων (χαρακτηριστικών)
 - Μείωση πλήθους δεδομένων

Συσσώρευση κύβων



- Το χαμηλότερο επίπεδο του κύβου
 - Τα αθροιστικά δεδομένα για μία ξεχωριστή οντότητα ενδιαφέροντος
 - Π.χ. ένας πελάτης σε μία ΑΔ τηλεφωνικής εταιρίας
- Πολλά επίπεδα συσσώρευσης μέσα στους κύβους
 - Περαιτέρω μείωση του μεγέθους των δεδομένων
- Αναφορά σε κατάλληλα επίπεδα
 - Χρήση της μικρότερης αναπαράστασης για την επίλυση δεδομένου προβλήματος

Μείωση διαστάσεων

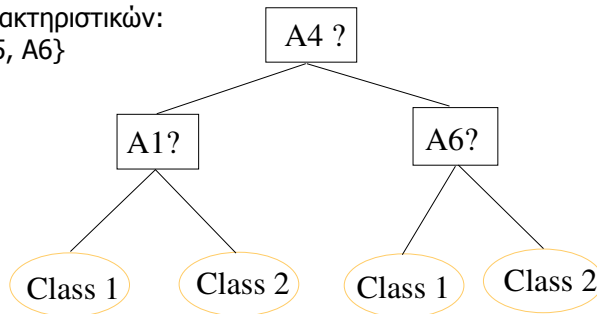


- Επιλογή χαρακτηριστικών (π.χ., επιλογή υποσυνόλου χαρακτηριστικών):
 - Επιλογή ελάχιστου συνόλου χαρακτηριστικών τέτοια ώστε η κατανομή πιθανότητας των διαφορετικών κλάσεων δεδομένων των τιμών για αυτά τα χαρακτηριστικά είναι το δυνατόν εγγύτερα στην αρχική κατανομή δοσμένων των τιμών όλων των χαρακτηριστικών
- Ευριστικές μέθοδοι:
 - Επαγωγή με δένδρα απόφασης (decision-tree induction)
 - Ανάλυση κυρίων συνιστωσών (principal components analysis – PCA)

Παράδειγμα επαγωγής με δέντρα απόφασης



- Αρχικό σύνολο χαρακτηριστικών: $\{A1, A2, A3, A4, A5, A6\}$



- Τα χαρακτηριστικά A2, A3, A5 δεν εμπλέκονται στο ΔΑ, άρα απαλείφονται → Μειωμένο σύνολο χαρακτηριστικών: $\{A1, A4, A6\}$

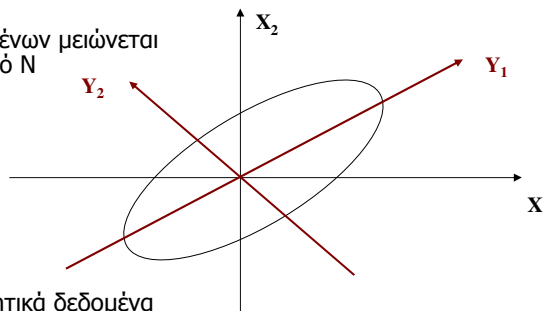
21

ΠΑ.ΠΕΛ – Γιάννης Θεοδορίδης

Ανάλυση κύριων συνιστωσών



- Principal Components Analysis (PCA)
- Πρόβλημα: Δοθέντων N διανυσμάτων δεδομένων k -διαστάσεων, να βρεθούν $c \leq k$ ορθογώνια διανύσματα που να αναπαριστούν βέλτιστα τα δεδομένα
 - Το αρχικό σύνολο δεδομένων μειώνεται σε ένα αποτελούμενο από N διανύσματα σε c κύριες συνιστώσες (μειωμένες διαστάσεις)
 - Κάθε διάνυσμα αποτελεί γραμμικό συνδυασμό των c διανυσμάτων κύριων συνιστωσών
 - Δουλεύει μόνο με αριθμητικά δεδομένα
 - Χρησιμεύει για μεγάλο αριθμό διαστάσεων



22

ΠΑ.ΠΕΛ – Γιάννης Θεοδορίδης

Μείωση πλήθους δεδομένων



- Παραμετρικές μέθοδοι
 - Υποθέτουμε ότι τα δεδομένα ακολουθούν κάποιο μοντέλο, υπολογίζουμε τις παραμέτρους του μοντέλου, αποθηκεύουμε μόνο τις παραμέτρους, και αποβάλλουμε τα δεδομένα (εκτός πιθανών outliers)
- Μη-παραμετρικές μέθοδοι
 - Δεν υποθέτουν μοντέλα
 - Βασικές τεχνικές: ιστογράμματα, συσταδοποίηση, δειγματοληψία

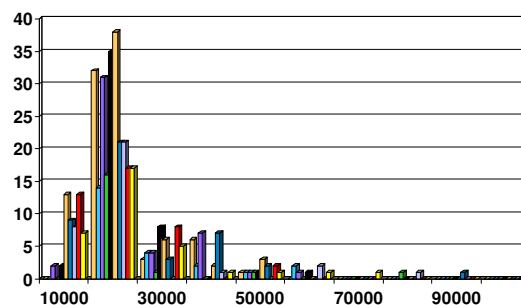
23

ΠΑ.ΠΕΛ – Γιάννης Θεοδορίδης

Ιστογράμματα



- Μία δημοφιλή τεχνική μείωσης δεδομένων
- Διαμερισμός των δεδομένων σε «κάδους» και αποθήκευση μέσου όρου κάθε «κάδου»



24

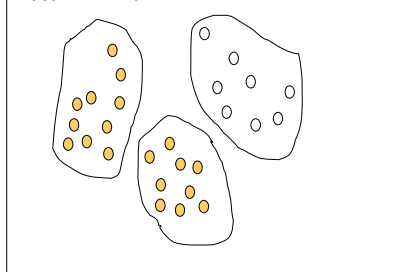
ΠΑ.ΠΕΛ – Γιάννης Θεοδορίδης

Δειγματοληψία



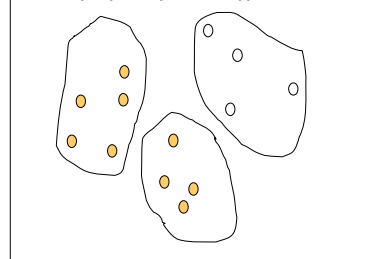
- Επιλογή ενός αντιπροσωπευτικού υποσυνόλου των δεδομένων
 - Η απλή τυχαία δειγματοληψία μπορεί να έχει χαμηλή απόδοση με την παρουσία ασύμμετρων δεδομένων
 - Διαστρωματωμένη δειγματοληψία:
 - Προσέγγιση του ποσοστού κάθε κλάσης (ή υπο-πληθυσμού ενδιαφέροντος) σε όλη τη ΒΔ

Αρχικά δεδομένα



25

Διαστρωματωμένο δείγμα



ΠΑ.ΠΕΛ – Γιάννης Θεοδορίδης

Εργασίες Προπαρασκευής Δεδομένων



- Καθαρισμός δεδομένων
 - Συμπλήρωση ελλειπών τιμών, εντοπισμός 'θορύβου', επίλυση ασυνεπειών
- Μετασχηματισμός δεδομένων
 - Εξομάλυνση, κανονικοποίηση τιμών
- Μείωση δεδομένων
 - Μείωση διαστάσεων (χαρακτηριστικών), μείωση πλήθους δεδομένων
- Διακριτοποίηση δεδομένων
 - Από συνεχείς σε διακριτές τιμές ή διαστήματα τιμών

26

ΠΑ.ΠΕΛ – Γιάννης Θεοδορίδης

Διακριτοποίηση



- Τρεις τύποι χαρακτηριστικών:
 - Ονομαστικά (nominal)-τιμές από μη διατεταγμένο σύνολο
 - χρώμα, όνομα, ...
 - Διατεταγμένα (Ordinal) - τιμές από διατεταγμένο σύνολο
 - βαθμολογία (Α – Β – Γ), ικανοποίηση (υψηλή – μέτρια – χαμηλή), ...
 - Συνεχή (Continuous) – πραγματικοί αριθμοί
 - βάρος, ύψος, μισθός, ηλικία, ...
- Διακριτοποίηση :
 - Διαχωρισμός του εύρους τιμών συνεχών χαρακτηριστικών σε διαστήματα
 - Γιατί; μείωση χώρου δεδομένων, προετοιμασία για περαιτέρω ανάλυση

27

ΠΑ.ΠΕΛ – Γιάννης Θεοδορίδης

Διακριτοποίηση για αριθμητικά δεδομένα



- Χρησιμοποιούνται μέθοδοι που έχουμε ήδη μελετήσει
 - Binning
 - Ιστογράμματα
 - Συσταδοποίηση
- ... αλλά και άλλες μέθοδοι
 - Διακριτοποίηση βασισμένη στην εντροπία
 - Κατάτμηση με φυσικό διαμερισμό

28

ΠΑ.ΠΕΛ – Γιάννης Θεοδορίδης

Διακριτοποίηση για ονομαστικά δεδομένα



- Καθορισμός μίας μερικής διάταξης χαρακτηριστικών σε επίπεδο σχήματος από χρήστες ή ειδικούς
- Καθορισμός ενός τμήματος μίας ιεραρχίας με ομαδοποίηση δεδομένων
- Καθορισμός ενός συνόλου χαρακτηριστικών, αλλά όχι της μερικής τους διάταξης
- Καθορισμός ενός μερικού συνόλου των δεδομένων

Σύνοψη



- Η προπαρασκευή δεδομένων είναι μεγάλο θέμα στο χώρο του Data Warehousing και Data Mining
- Περιλαμβάνει καθαρισμό, μετασχηματισμό, μείωση δεδομένων
 - Συμπλήρωση ελλειπών τιμών, εντοπισμός 'θορύβου', επίλυση ασυνεπειών
 - Εξομάλυνση, κανονικοποίηση τιμών
 - Μείωση διαστάσεων (χαρακτηριστικών), μείωση πλήθους δεδομένων
 - Διακριτοποίηση δεδομένων