2

Επεξεργασία Δεδομένων

Περιεχόμενα Κεφαλαίου

2.1 Μορφή και Ιδιότητες Δεδομένων

Ως *σύνολο δεδομένων* εννοούμε μία δεδομένη συλλογή αντικειμένων. Κάθε αντικείμενο έχει τη μορφή μίας εγγραφής, που αποτελείται από ένα σύνολο πεδίων. Τα πεδία καλούνται *ιδιότητες* (attributes) ή *χαρακτηριστικά* (features). Εφεξής, οι δύο αυτές ονομασίες θα χρησιμοποιούνται ισοδύναμα. Συνήθως, η αναπαράσταση ενός συνόλου δεδομένων γίνεται με τη μορφή ενός πίνακα, όπως περιγράφεται στο ακόλουθο παράδειγμα.

Παράδειγμα 2.1.1 (Αναπαράσταση πίνακα) Υποθέτουμε ένα σύνολο δεδομένων, που αφορά σε βιβλία εκμάθησης γλωσσών Προγραμματισμού, τα οποία πωλούνται από ένα ηλεκτρονικό βιβλιοπωλείο. Κάθε αντικείμενο (βιβλίο) του συνόλου αυτού έχει τις εξής ιδιότητες: {*κωδικός, τίτλος, σελίδες, επίπεδο, τιμή, διαθεσιμότητα*}. Ο Πίνακας 2.1 περιέχει κάποια βιβλία και ενδεικτικές τιμές για τις ιδιότητές τους. □

κωδικός	*τίτλος*	*σελίδες*	*επίπεδο*	*τιμή*	*διαθέσιμο*
525745	Εισαγωγή στη Java	256	Χαμηλό	57.20	Ναι
525746	Προχωρημένη C++	427	Υψηλό	64.50	Όχι
525747	Εφαρμογές στην SQL	357	Μεσαίο	45.30	Ναι

Πίνακας 2.1: Αναπαράσταση συνόλου δεδομένων με μορφή πίνακα.

Κάθε ιδιότητα μπορεί να θεωρηθεί ως μεταβλητή που λαμβάνει τιμές από ένα πεδίο ορισμού. Αναλόγως του πεδίου ορισμού, οι ιδιότητες μπορεί να ανήκουν σε μία από τις εξής δύο κατηγορίες:

Διακριτές. Όταν το πεδίο ορισμού μίας ιδιότητας είναι ένα πεπερασμένο, ή μη πεπερασμένο αλλά αριθμήσιμο σύνολο διακριτών αντικειμένων, τότε η ιδιότητα ονομάζεται *διακριτή* (discrete). Αν δεν ορίζεται μία σχέση διάταξης στο πεδίο ορισμού της, τότε η ιδιότητα καλείται *ονομαστική* (nominal), αλλιώς καλείται *διατεταγμένη* (ordinal). Οι *δυαδικές* (binary) ιδιότητες είναι μία ειδική περίπτωση ονομαστικών ιδιοτήτων, όπου το πεδίο ορισμού αποτελείται από δύο τιμές, όπως για παράδειγμα οι περιπτώσεις: {*true, false*}, {*yes, no*}, {*0, 1*}, κ.λπ.

Συνεχείς. Μία ιδιότητα καλείται *συνεχής* (continuous), αν το πεδίο ορισμού της είναι μη πεπερασμένο και μη αριθμήσιμο, και επιπλέον ορίζεται και μία σχέση διάταξης επί των τιμών της. Συνήθως, μία ιδιότητα είναι συνεχής αν οι τιμές της είναι πραγματικοί αριθμοί.

Στον Πίνακα 2.1, διακριτές είναι οι ιδιότητες *κωδικός, τίτλος, σελίδες, επίπεδο*, και *διαθέσιμο*, ενώ η ιδιότητα *τιμή* είναι συνεχής. Επιπλέον, οι ιδιότητες *κωδικός, τίτλος*, και *διαθέσιμο* είναι ονομαστικές, επειδή δεν ορίζεται κάποια σχέση διάταξης επ' αυτών. Αντιθέτως, οι ιδιότητες *σελίδες* και *επίπεδο* είναι διατεταγμένες.

Μία διαφορετική κατηγοριοποίηση των ιδιοτήτων γίνεται αναλόγως: (α) αν το πεδίο ορισμού είναι το σύνολο των ακεραίων ή των πραγματικών αριθμών, ή (β) αν περιέχει αλφαριθμητικά. Στην πρώτη περίπτωση η ιδιότητα ονομάζεται *αριθμητική* (numeric), ενώ στη δεύτερη ονομάζεται *συμβολική* (symbolic). Παρατηρούμε ότι μία αριθμητική ιδιότητα μπορεί να είναι διακριτή (αν παίρνει ακέραιες τιμές) ή συνεχής (αν παίρνει πραγματικές τιμές).

Στον Πίνακα 2.1, οι ιδιότητες *σελίδες* και *τιμή* είναι αριθμητικές. Αντιθέτως, οι ιδιότητες *κωδικός, τίτλος, επίπεδο* και *διαθέσιμο* είναι συμβολικές. Αν και φαίνεται ότι η ιδιότητα *κωδικός* λαμβάνει ακέραιες τιμές, εντούτοις παρατηρούμε ότι το πεδίο ορισμού της αποτελεί απλώς μία αντιστοίχηση 1-1, επί της οποίας δεν ορίζονται πράξεις όπως η διάταξη, η πρόσθεση, κ.λπ. Επομένως, η σημασία μίας ιδιότητας καθορίζει τελικώς την κατηγορία όπου αυτή ανήκει.

Στην ειδική περίπτωση που όλες οι ιδιότητες ενός συνόλου δεδομένων είναι αριθμητικές, τότε κάθε αντικείμενό του ονομάζεται *διάνυσμα* (ή, εναλλακτικά, *σημείο*), ενώ οι ιδιότητές του ονομάζονται διαστάσεις. Ο αριθμός των διαστάσεων ονομάζεται *διαστασιμότητα* (dimensionality).

2.2 Περίληψη Δεδομένων

Ως περίληψη δεδομένων καλούμε τη δυνατότητα δημιουργίας μίας συνολικής εικόνας των ιδιοτήτων των δεδομένων. Μέτρα και διαγράμματα για αυτό το σκοπό είναι τα εξής:

Μέση τιμή (mean value). Έστω οι αριθμητικές τιμές $x_1, \ldots, x_n$. Η μέση τιμή τους ορίζεται ως $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$. Παράδειγμα: η μέση τιμή της ακολουθίας 5, 3, 4, 9, 6, 5, 7 είναι ίση με 5.57.

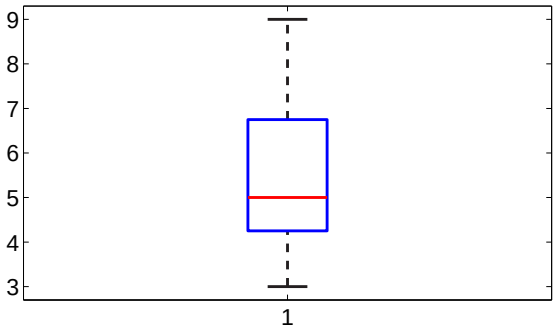
Ενδιάμεση τιμή (mean value). Η ενδιάμεση τιμή είναι αντιπροσωπευτικότερη της μέσης για ακολουθία με ανομοιόμορφη κατανομή τιμών. Αν η ακολουθία έχει περιττό πλήθος στοιχείων, τότε ως η ενδιάμεση τιμή ορίζεται η τιμή που βρίσκεται στη μέση της ακολουθίας μετά την αύξουσα ταξινόμησή της. Παράδειγμα: η ενδιάμεση τιμή της ακολουθίας 5, 3, 4, 9, 6, 5, 7 είναι ίση με 5. Αν το πλήθος των στοιχείων είναι άρτιο, τότε ως ενδιάμεση τιμή λαμβάνεται το ημιάθροισμα των δύο μεσαίων όρων μετά την αύξουσα

ταξινόμησή της. Παράδειγμα: η ενδιάμεση τιμή της ακολουθίας 3, 4, 9, 6, 5, 7 είναι ίση με 5.5.

Mode. Ως mode τιμή μίας ακολουθίας ορίζουμε την τιμή με τη μεγαλύτερη συχνότητα εμφάνισης. Παράδειγμα: η mode τιμή της ακολουθίας 5, 3, 4, 9, 6, 5, 7 είναι ίση με 5.

Διασπορά (variance). Η διασπορά μίας ακολουθίας τιμών $x_1, \ldots, x_n$ ορίζεται ως $Var(x) = \mu = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$. Παράδειγμα: η διασπορά της ακολουθίας 5, 3, 4, 9, 6, 5, 7 είναι ίση με 3.95.

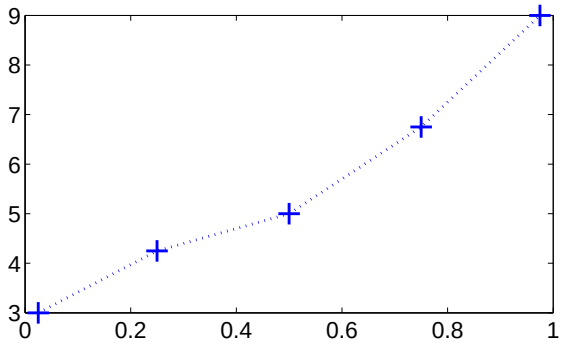
Διάγραμμα boxplot. Η περίληψη των δεδομένων εμφανίζεται μέσω ενός διαγράμματος boxplot. Σε αυτό απεικονίζονται 5 ποσότητες: η ελάχιστη και η μέγιστη τιμή στα δύο άκρα του, ενώ τα όρια του 'κουτιού' απεικονίζουν το πρώτο τεταρτημόριο (η 25% μεγαλύτερη τιμή) και το τρίτο τεταρτημόριο (η 75% μεγαλύτερη τιμή). Εντός του 'κουτιού' απεικονίζεται η ενδιάμεση τιμή. Παράδειγμα: το διάγραμμα boxplot της ακολουθίας 5, 3, 4, 9, 6, 5, 7 απεικονίζεται στο Σχήμα 2.1.



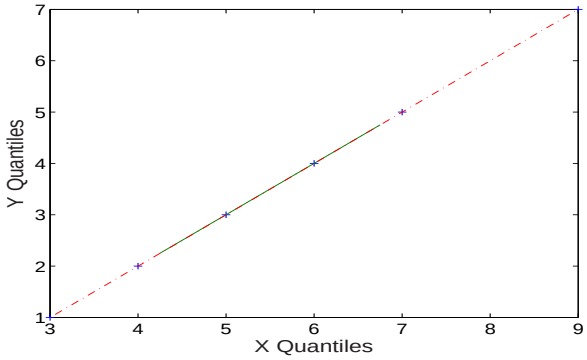
Σχήμα 2.1: Διάγραμμα boxplot.

Διαγράμματα quantile plot και qqplot. Τα διαγράμματα quantile plot αποτυπώνουν τις τιμές των διαφόρων τεταρτημορίων μίας ακολουθίας. Παράδειγμα: το διάγραμμα quantile plot της ακολουθίας 5, 3, 4, 9, 6, 5, 7 για τα τεταρτημόρια 2.5%, 25%, 50%, 75%, και 97.5% απεικονίζεται στο Σχήμα 2.2. Ουσιαστικά δηλώνει το ποσοστό των αντικειμένων με τιμή μικρότερη από μία συγκεκριμένη.

Ένα διάγραμμα qqplot προβάλλει συγκριτικά δύο quantile plots δύο διαφορετικών ακολουθιών και χρησιμοποιείται για τη σύγκρισή τους. Παρά-



Σχήμα 2.2: Διάγραμμα quantile plot.



Σχήμα 2.3: Διάγραμμα qqplot.

δειγμα: το διάγραμμα quantile plot της ακολουθίας X =5, 3, 4, 9, 6, 5, 7 για τα τεταρτημόρια 2.5%, 25%, 50%, 75%, και 97.5% σε σύγκριση με τα αντίστοιχα της ακολουθίας Y =3, 1, 2, 7, 4, 3, 5, απεικονίζεται στο Σχήμα 2.3. Στο qqplot διάγραμμα ταξινομούμε τις δύο ακολουθίες και κάθε ζεύγος τιμών τους αναπαρίσταται με ένα σημείο (απεικονίζεται με το σύμβολο '+') στις δύο διαστάσεις. Αν τα δείγματα προέρχονται από την ίδια κατανομή (στο παράδειγμα αυτό ισχύει, επειδή $Y = X - 2$), τότε το διάγραμμα είναι γραμμικό. Η γραμμή συνδέει το πρώτο με το τρίτο τεταρτημόριο και είναι για εποπτικούς λόγους, δηλαδή για τον έλεγχο της γραμμικότητας.

2.3 Καθαρισμός και Μετασχηματισμός Δεδομένων

Στις περισσότερες εφαρμογές εξόρυξης δεδομένων, τα δεδομένα προϋπάρχουν και έχουν συλλεχθεί με διαδικασίες που δεν είχαν σχεδιασθεί ειδικά για να διευκολύνουν το έργο της εξόρυξής τους. Επιπλέον, σχεδόν σε όλες τις διαδικασίες συλλογής πραγματικών δεδομένων ενυπάρχουν παράγοντες, όπως ανθρώπινα λάθη, προβλήματα με συσκευές μέτρησης, κακός σχεδιασμός διαδικασιών κ.ο.κ. Το αποτέλεσμα όλων αυτών είναι η *ποιότητα των δεδομένων* να μην είναι πάντοτε η καλύτερη δυνατή. Για παράδειγμα, η ποιότητα των δεδομένων επηρεάζεται από την ύπαρξη:

- Θορύβου στα δεδομένα (από ανθρώπινα λάθη ή προβλήματα συσκευών).
- "Ανώμαλων" (outliers) και ασυνεπών τιμών (από λάθη κατά την εισαγωγή και κακό σχεδιασμό).
- Ελλιπών τιμών (από κακό σχεδιασμό ή λάθη κατά την εισαγωγή).

Αν τα προαναφερθέντα προβλήματα δεν αντιμετωπισθούν εξ αρχής, τότε είτε δεν θα μπορεί να εφαρμοσθεί η εξόρυξη δεδομένων, είτε θα αλλοιωθεί σημαντικά η ποιότητα των αποτελεσμάτων (σύμφωνα με τη γνωστή αρχή του garbage in, garbage out). Για την αντιμετώπιση των προβλημάτων αυτών, χρησιμοποιούνται τεχνικές καθαρισμού των δεδομένων. Επειδή η επιλογή των τεχνικών καθαρισμού σχετίζεται με τα προβλήματα που παρατηρούνται σε κάθε περίπτωση ξεχωριστά, δεν υπάρχουν γενικές προσεγγίσεις. Στη συνέχεια, θα περιγράψουμε κάποια στοιχεία τεχνικών καθαρισμού, αποσκοπώντας στην κατανόηση των βασικών εννοιών.

Ανεξαρτήτως του καθαρισμού των δεδομένων, συχνά είναι απαραίτητος ο μετασχηματισμός των δεδομένων. Για παράδειγμα, αν έχουμε μεταβλητές με πολύ διαφορετικά διαστήματα τιμών, ή αν ο αλγόριθμος απαιτεί τα δεδομένα εισόδου να ανήκουν σε ένα συγκεκριμένο διάστημα, τότε εφαρμόζουμε κάποιο μετασχηματισμό επί των τιμών των δεδομένων. Όπως για τον καθαρισμό, έτσι και για το μετασχηματισμό δεδομένων δεν υπάρχουν γενικές τεχνικές. Στη συνέχεια θα περιγράψουμε μόνο κάποιους βασικούς αλλά συχνά χρησιμοποιούμενους μετασχηματισμούς.

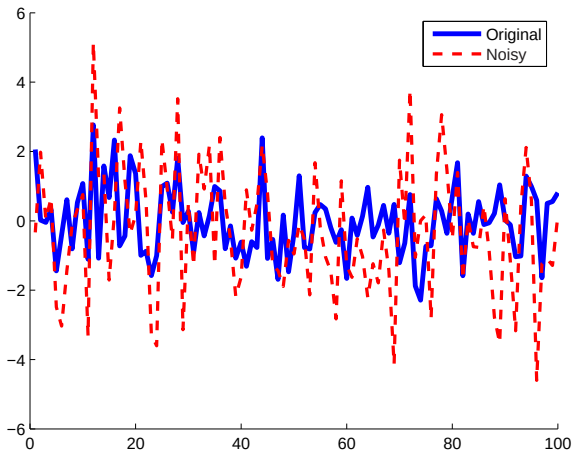
Η επιτυχής εφαρμογή τεχνικών καθαρισμού και μετασχηματισμού ενισχύεται από τη γνώση μας για τα ίδια τα δεδομένα. Η ύπαρξη τεκμηρίωσης, π.χ., με τη μορφή *μετα-δεδομένων*, για τον τρόπο συλλογής τους, για την αναπαράστασή τους και τη σημασία τους, βοηθά όσο τίποτε άλλο. Επίσης, πρέπει να γνωρίζουμε ότι όσο καλά και αν καθαρίσουμε τα δεδομένα, ίσως να μην μπορούμε

να εξαλείψουμε την ύπαρξη προβλημάτων. Για αυτόν το λόγο είναι απαραίτητο να έχουμε αλγορίθμους εξόρυξης, οι οποίοι να είναι ανεκτικοί σε προβληματικά δεδομένα.

2.3.1 Καθαρισμός δεδομένων

Θόρυβος

Θόρυβο (noise) καλούμε είτε την τυχαία αλλοίωση τιμών είτε την παρείσφρηση αντικειμένων με τυχαίες τιμές. Ενδεχόμενο αποτέλεσμα είναι τα θορυβώδη αντικείμενα να μην ακολουθούν τα πρότυπα των υπολοίπων, δυσχεραίνοντας το έργο της εξόρυξής τους. Συνήθως, για την αφαίρεση του θορύβου ακολουθούνται τεχνικές επεξεργασίας σήματος (ιδιαίτερα για μεταβλητές με συνεχείς τιμές). Το ακόλουθο παράδειγμα περιγράφει μία τέτοια τεχνική, που βασίζεται στην τεχνική της *εξομάλυνσης κυλιόμενου μέσου όρου* (moving-average smoothing).



Σχήμα 2.4: Σειρά 100 σημείων και θορυβώδης σειρά.

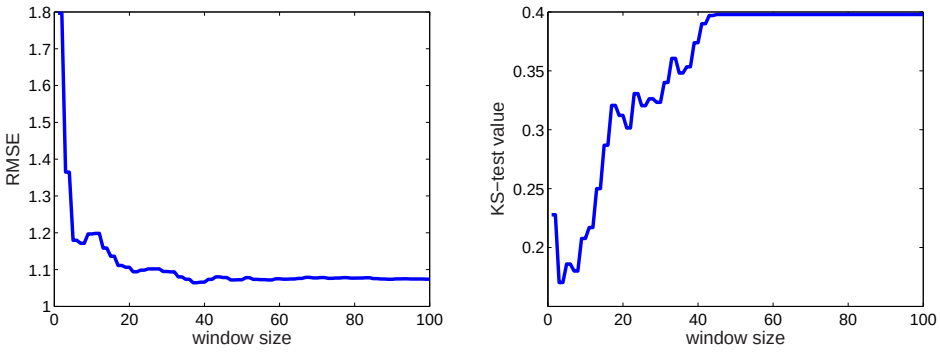
Παράδειγμα 2.3.1 (Εξομάλυνση κυλιόμενου μέσου όρου) Στο Σχήμα 2.4 με συνεχή γραμμή απεικονίζεται μια σειρά 100 τιμών που ακολουθούν τυποποιημένη κανονική κατανομή (μέση τιμή 0 και τυπική απόκλιση 1). Στο ίδιο σχήμα με διακεκομμένη γραμμή απεικονίζεται η ίδια σειρά, όπου όμως οι τιμές έχουν

αλλοιωθεί με λευκό θόρυβο.[1] Η συνολική αλλοίωση συνήθως μετράται με τη *ρίζα του μέσου τετραγωνικού λάθους* (root mean square error). Αν $X = \langle x_1, \ldots, x_n \rangle$ είναι η αρχική σειρά και $Y = \langle y_1, \ldots, y_n \rangle$ η θορυβώδης, τότε:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{x_i} - \mathrm{y_i})^2}{\mathrm{n}}}$$

Για τα δεδομένα του Σχήματος 2.4 προκύπτει ότι RMSE=1.8.

Εφαρμόζουμε την τεχνική του κυλιόμενου μέσου, όπου θεωρούμε ένα παράθυρο με μήκος w. Θέτουμε κάθε τιμή της θορυβώδους σειράς ίση με το μέσο όρο των $w/2$ τιμών στα αριστερά της και των $w/2$ των τιμών στα δεξιά της.[2] Δοκιμάζουμε διάφορες τιμές για το w και κάθε φορά μετρούμε το RMSE για κάθε σειρά που προκύπτει αντιστοίχως. Τα αποτελέσματα απεικονίζονται στο Σχήμα 2.5α. Προκύπτει ότι μεγαλύτερα μήκη παραθύρου w μειώνουν το RMSE. Όμως, καθώς η τιμή του w μεγαλώνει, η σειρά που προκύπτει τείνει να είναι σταθερή και ίση με την τιμή του συνολικού μέσου όρου. Επομένως, το μέτρο RMSE δεν είναι ενδεικτικό του κατά πόσο η εξομαλυμένη σειρά ταιριάζει με την αρχική.



(α) (β)

Σχήμα 2.5: (α) RMSE ως προς μήκος παραθύρου w, (β) Τιμή Kolmogorov-Smirnov ως προς μήκος παραθύρου w.

Καθώς είναι γνωστό ότι οι τιμές της αρχικής σειράς ακολουθούν μία τυποποιημένη κανονική κατανομή, θα εφαρμόσουμε τον έλεγχο Kolmogorov-

[1] Ως λευκό θόρυβο ονομάζουμε μία σειρά με τιμές που ακολουθούν κανονική κατανομή και προστίθενται στη δεδομένη σειρά.

[2] Αν δεν είναι δυνατόν να λάβουμε $w/2$ τιμές προς κάποια κατεύθυνση, τότε το μήκος του παραθύρου περικόπτεται αναλόγως.

Smirnov, σύμφωνα με τον οποίο ελέγχεται αν δύο κατανομές διαφέρουν ή αν μία δεδομένη κατανομή διαφέρει από μία υποθετική. Γενικώς, ο έλεγχος Kolmogorov-Smirnov επιστρέφει μία τιμή στο διάστημα [0,1], όπου όσο μικρότερη είναι η τιμή αυτή, τόσο πιθανότερο είναι οι τιμές της σειράς να ακολουθούν την τυποποιημένη κανονική κατανομή. Τα αποτελέσματα απεικονίζονται στο Σχήμα 2.5β. Όταν το w ισούται με 5, τότε η εξομαλυμένη σειρά έχει τη μεγαλύτερη πιθανότητα να ακολουθεί την τυποποιημένη κανονική κατανομή. Αντιθέτως, καθώς το w μεγαλώνει, τόσο αυτή η πιθανότητα μικραίνει σημαντικά. Άρα, το αποτέλεσμα της εξομάλυνσης είναι καλύτερο για σχετικά μικρές τιμές παραθύρου. □

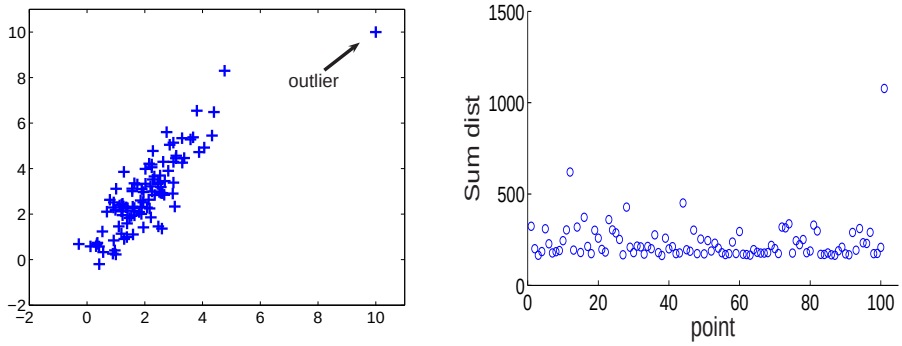
Ανώμαλες και ασυνεπείς τιμές

Αν οι τιμές κάποιων ιδιοτήτων ενός αντικειμένου δεν ακολουθούν την κατανομή των τιμών των υπολοίπων αντικειμένων, τότε τις αποκαλούμε *ανώμαλες*, ενώ τα αντικείμενα με ανώμαλες τιμές καλούνται *outliers*. Συνήθως τέτοιες τιμές προκύπτουν από λάθη κατά την εισαγωγή τιμών. Όμως σε κάποιες περιπτώσεις μπορεί να μην υπάρχει λάθος, αλλά η διαφοροποίηση να δηλώνει ένα ενδιαφέρον φαινόμενο (π.χ., την περίπτωση μίας οικονομικής απάτης, όπως θα αναλυθεί περισσότερο σε επόμενο κεφάλαιο). Στη συνέχεια, περιγράφουμε ένα παράδειγμα μίας απλής τεχνικής εντοπισμού αντικειμένων με ανώμαλες τιμές σε αριθμητικές ιδιότητες.

Παράδειγμα 2.3.2 (Εντοπισμός ανώμαλης τιμής) Στο Σχήμα 2.6α απεικονίζονται οι συντεταγμένες 100 σημείων που ακολουθούν δισδιάστατη κανονική κατανομή, και ένα επιπλέον σημείο (το εκατοστό πρώτο, με την ένδειξη outlier) που δεν ακολουθεί την κατανομή των υπολοίπων.

Μετρούμε το άθροισμα της απόσταση κάθε σημείου από όλα τα υπόλοιπα. Ως μέτρο απόστασης θεωρούμε την Ευκλείδια απόσταση. Στο Σχήμα 2.6β για κάθε σημείο απεικονίζεται το άθροισμα των αποστάσεων από όλα τα υπόλοιπα. Προκύπτει ότι το σημείο που δεν ακολουθεί την κατανομή των υπολοίπων (δηλαδή, το εκατοστό πρώτο) απέχει κατά πολύ περισσότερο από τα υπόλοιπα. Εφόσον διαπιστώσουμε ότι πρόκειται για λάθος, μπορούμε να διορθώσουμε τις τιμές του εντοπισμένου σημείου. □

Ως *ασυνεπείς* (inconsistent) καλούμε τις τιμές που δεν έχουν νόημα. Παραδείγματα είναι: βάρος ανθρώπου 1025 κιλά, ημερομηνία γέννησης 10982 μ.Χ., ταχυδρομικός κωδικός -20355, κ.λπ. Η συνέπεια μίας τιμής ορίζεται στα πλαίσια της εκάστοτε εφαρμογής. Ως εκ τούτου, εφόσον σχεδιασθεί σωστά μία



(α) (β)

Σχήμα 2.6: (α) Απεικόνιση σημείων, (β) Συνολική απόσταση κάθε σημείου από τα υπόλοιπα.

εφαρμογή και ορισθούν τα έγκυρα πεδία ορισμού των ιδιοτήτων της, κατά την εισαγωγή των τιμών μπορεί να ελέγχεται αυτόματα αν παραβιάζεται κάποιο πεδίο ορισμού.

Ελλιπείς τιμές

Σε πραγματικά σύνολα δεδομένων, είναι συχνό φαινόμενο κάποια αντικείμενα να περιέχουν *ελλιπείς τιμές* (missing values), δηλαδή χαρακτηριστικά με άγνωστη τιμή. Ελλιπείς τιμές προκύπτουν εξαιτίας διαφόρων αιτιών. Για παράδειγμα, σε έρευνα για συλλογή στατιστικών στοιχείων, κάποια δεδομένα θεωρούνται προσωπικά, οπότε δεν μπορούμε να τα έχουμε στη διάθεσή μας για το σύνολο των ερωτηθέντων. Επίσης, σε δεδομένα που συλλέγονται από ερωτηματολόγια στον παγκόσμιο ιστό, κάποιοι χρήστες μπορεί να μην συμπληρώνουν τα προαιρετικά πεδία της φόρμας, κ.λπ.

Για την αντιμετώπιση του προβλήματος των ελλιπών τιμών, θεωρούμε τις εξής προσεγγίσεις:

- Διαγραφή αντικειμένων με ελλιπείς τιμές. Αν ο αριθμός των αντικειμένων με ελλιπείς τιμές είναι μεγάλος, τότε μπορεί να μην είναι αντιπροσωπευτικό το σύνολο δεδομένων που προκύπτει.

- Συμπλήρωση ελλιπών τιμών μέσω εκτίμησης. Η συμπλήρωση μπορεί να γίνει με διάφορες τεχνικές εκτίμησης, οι οποίες αντιμετωπίζουν το πρόβλημα εφόσον είναι καλές. Διαφορετικά, οι ελλιπείς τιμές μετατρέπονται

σε θορυβώδεις. Συχνά χρησιμοποιούμενες τεχνικές είναι η συμπλήρωση με το μέσο όρο της τιμής, μέσω παλινδρόμησης, κ.α.

- Διατήρηση ελλιπών τιμών. Εφόσον ο αλγόριθμος εξόρυξης μπορεί να λειτουργήσει με ελλιπείς τιμές (είτε απλώς αγνοώντας τες είτε συμπληρώνοντάς τες), τότε δεν κάνουμε κάποια ενέργεια αντιμετώπισης των ελλιπών τιμών. Όμως πρέπει να σημειωθεί ότι η ποιότητα του αποτελέσματος του αλγορίθμου σχεδόν πάντα επηρεάζεται από την ύπαρξη ελλιπών τιμών.

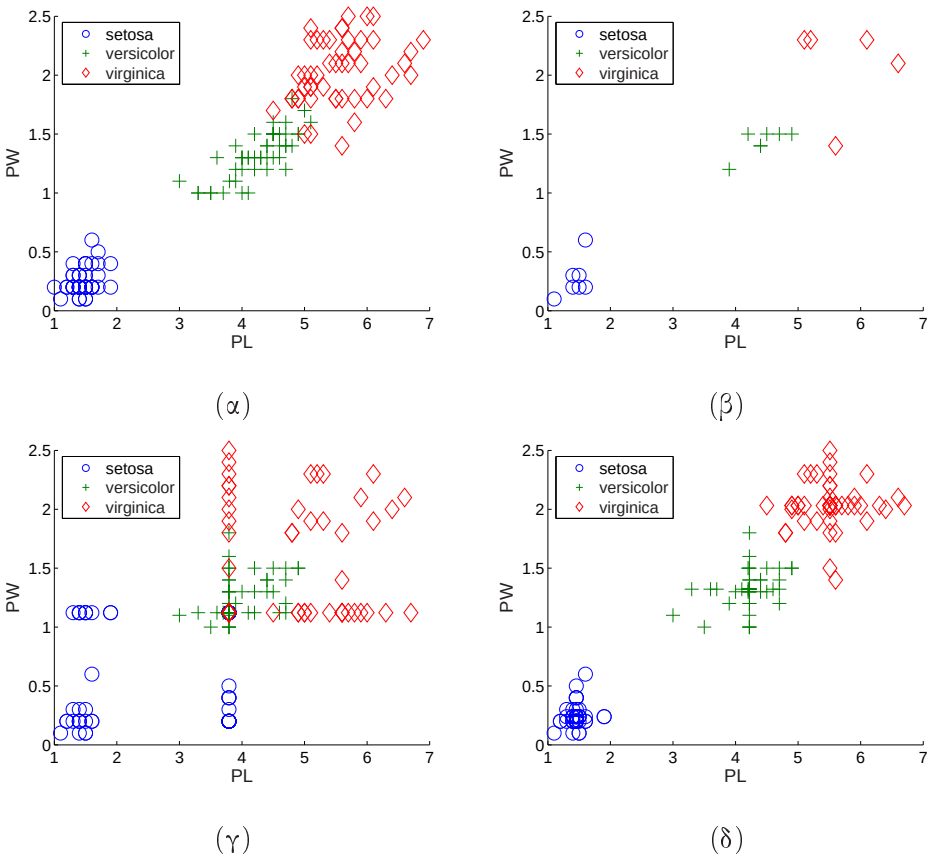
Το επόμενο παράδειγμα αποτυπώνει τις προαναφερθείσες προσεγγίσεις.

Παράδειγμα 2.3.3 (Αντιμετώπιση ελλιπών τιμών) Για παράδειγμα, χρησιμοποιούμε το σύνολο δεδομένων Iris, το οποίο περιέχει 150 λουλούδια που χωρίζονται σε τρεις κλάσεις (είδη): Setosa, Versicolor, Virginica. Για κάθε λουλούδι μετρούμε τις τιμές τεσσάρων αριθμητικών χαρακτηριστικών: μήκος σέπαλου (SL), πλάτος σέπαλου (SW), μήκος πετάλου (PL), πλάτος πετάλου (PW). Στο Σχήμα 2.7α απεικονίζονται τα σημεία του συνόλου Iris, επιλέγοντας για διαστάσεις τα δύο τελευταία χαρακτηριστικά (το θέμα της επιλογής χαρακτηριστικών θα αναλυθεί στο Κεφάλαιο 2.6.2).

Συνολικά, στο Iris περιέχονται $150 \times 4 = 600$ αριθμητικές τιμές. Επιλέγουμε τυχαία και διαγράφουμε το 60% από αυτές, μετατρέποντάς τες σε ελλιπείς τιμές. Αν και διαγράψαμε το 60%, εντούτοις προκύπτει ότι 131 από τα 150 αντικείμενα (δηλαδή, το 87.3%) έχουν τουλάχιστον ένα χαρακτηριστικό με ελλιπή τιμή.

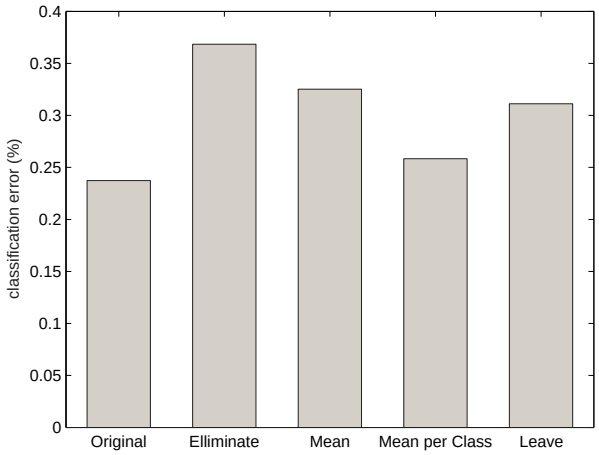
Αν θεωρήσουμε την προσέγγιση της διαγραφής των αντικειμένων με ελλιπείς τιμές, τότε παραμένουν μόνο 12.7% των αντικειμένων που απεικονίζονται στο Σχήμα 2.7β. Εξαιτίας του μεγάλου αριθμού διαγραμμένων αντικειμένων, είναι προφανές ότι αυτά που απομένουν δεν είναι αντιπροσωπευτικά του αρχικού συνόλου.

Στη συνέχεια, εξετάζουμε δύο εναλλακτικές μεθόδους εκτίμησης των ελλιπών τιμών με σκοπό τη συμπλήρωσή τους. Σύμφωνα με την πρώτη μέθοδο, οι ελλιπείς τιμές ενός χαρακτηριστικού συμπληρώνονται με το μέσο όρο των διαθέσιμων τιμών για το χαρακτηριστικό αυτό. Το αποτέλεσμα αυτό αποτυπώνεται στο Σχήμα 2.7γ, όπου παρατηρούμε ότι αν και τα αντικείμενα είναι περισσότερα από ότι στην προηγούμενη προσέγγιση, όμως πολλά αντικείμενα έχουν την ίδια τεχνητή τιμή για κάθε χαρακτηριστικό. Επομένως, το αποτέλεσμα και πάλι δεν είναι αντιπροσωπευτικό του αρχικού συνόλου. Σύμφωνα με τη δεύτερη εναλλακτική μέθοδο, οι ελλιπείς τιμές ενός χαρακτηριστικού, συμπληρώνονται με το μέσο όρο των διαθέσιμων τιμών για το χαρακτηριστικό αυτό μόνο μεταξύ των αντικειμένων της ίδιας κλάσης. Στο Σχήμα 2.7δ απεικονίζεται το αποτέλεσμα που ταιριάζει περισσότερο με το αρχικό σύνολο.



Σχήμα 2.7: (α) Σύνολο Iris, (β) Διαγραφή αντικειμένων με ελλιπείς τιμές, (γ) Συμπλήρωση μέσου όρου, (δ) Συμπλήρωση μέσου όρου ανά κλάση.

Το αποτέλεσμα των προσεγγίσεων ελέγχεται μέσω ενός αλγορίθμου κατηγοριοποίησης (η κατηγοριοποίηση αναλύεται στο Κεφάλαιο 6), δηλαδή υπολογίζοντας την ακρίβεια (=το ποσοστιαίο λάθος) που προκύπτει για κάθε προσέγγιση. Στο Σχήμα 2.8 παρουσιάζονται τα αποτελέσματα της ακρίβειας για πέντε δοκιμές, όπου η πρώτη τιμή αντιστοιχεί στο αρχικό σύνολο πριν την αλλοίωσή του, ενώ η τελευταία τιμή αντιστοιχεί στην περίπτωση όπου αποδεχόμαστε τις ελλιπείς τιμές. Από το σχήμα προκύπτει ότι η διατήρηση των ελλιπών τιμών ενδείκνυται μόνο όταν ο αλγόριθμος είναι σε θέση να τις αντιμετωπίσει ικανοποιητικά. □



Σχήμα 2.8: Λάθος κατηγοριοποίησης ανά προσέγγιση αντιμετώπισης ελλιπών τιμών.

2.3.2 Μετασχηματισμός δεδομένων

Μετασχηματισμό (transformation) δεδομένων ονομάζουμε την εφαρμογή κάποιας συνάρτησης επί των τιμών μίας ιδιότητας. Γενικώς, κάθε ιδιότητα μετασχηματίζεται ξεχωριστά, ενώ συνήθως επιλέγουμε για μετασχηματισμό μία αριθμητική ιδιότητα με πολύ μεγαλύτερες ή μικρότερες τιμές από τις τιμές των υπολοίπων ιδιοτήτων. Ο λόγος είναι ότι αυτή η διαφοροποίηση μπορεί να πολώσει αρνητικά τα αποτελέσματα ενός αλγορίθμου εξόρυξης. Ένας συχνά χρησιμοποιούμενος μετασχηματισμός για αυτήν την περίπτωση είναι η *τυποποίηση* (standardization ή z-score). Αν $\bar{X}$ είναι η μέση τιμή των τιμών μίας ιδιότητας X και S η τυπική τους απόκλιση, τότε η τυποποίηση γίνεται με βάση το μετασχηματισμό:

$$X = \frac{X - \bar{X}}{S}$$

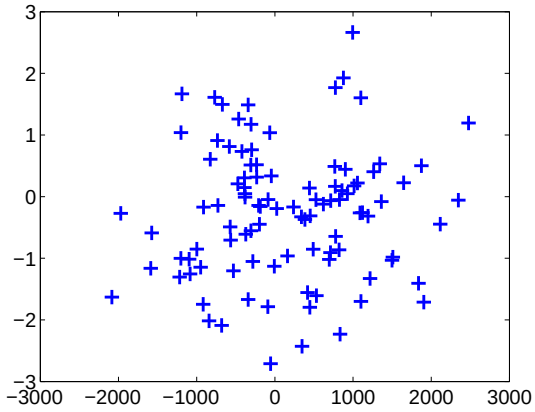
Άλλη συχνή περίπτωση χρήσης μετασχηματισμού είναι αν ο αλγόριθμος απαιτεί τα δεδομένα εισόδου να ανήκουν σε ένα συγκεκριμένο διάστημα, π.χ., από 0 έως 1. Έστω μία ιδιότητα X όπου $X_{\min}, X_{\max}$ είναι αντιστοίχως η ελάχιστη και μέγιστη τιμή της X. Αν θέλουμε η X να παίρνει τιμές εντός ενός νέου διαστήματος με ελάχιστη τιμή $X'_{\min}$ και μέγιστη τιμή $X'_{\max}$, τότε ο μετασχημα-

τισμός (που καλείται και min-max) γίνεται ως εξής:

$$X = \frac{X - X_{\min}}{X_{\max} - X_{\min}}(X'_{\max} - X'_{\min}) + X'_{\min}$$

Παράδειγμα 2.3.4 (Τυποποίηση και μετασχηματισμός διαστήματος) Στο Σχήμα 2.9 απεικονίζονται οι συντεταγμένες 100 δισδιάστατων σημείων που ακολουθούν κανονική κατανομή.

Η ιδιότητα που αντιστοιχεί στον οριζόντιο άξονα έχει εύρος τιμών ίσο με 5000, ενώ η άλλη ιδιότητα έχει εύρος τιμών 5. Υπολογίζουμε για κάθε δυνατό ζεύγος σημείων την Ευκλείδεια απόστασή τους, και στη συνέχεια λαμβάνουμε το μέσο όρο των αποστάσεων αυτών, ο οποίος προκύπτει ίσος με 987.26. Αν εκτελέσουμε την ίδια πράξη ξεχωριστά για κάθε διάσταση, τότε ο μέσος όρος αποστάσεων για την πρώτη ιδιότητα ισούται με 987.25, ενώ για την άλλη ιδιότητα με 1.08. Συνεπώς, η ιδιότητα με το μεγάλο εύρος κυριαρχεί κατά τη μέτρηση της Ευκλείδειας απόστασης μεταξύ των σημείων. Αν τυποποιήσουμε και τις δύο διαστάσεις, τότε ο μέσος όρος των ανά δύο αποστάσεων είναι ίσος με 1.79. Οι αντίστοιχοι μέσοι όροι για τις δύο διαστάσεις ξεχωριστά είναι ίσοι με 1.13 και 1.14. Άρα, πλέον δεν παρουσιάζεται το μειονέκτημα κάποια διάσταση να κυριαρχεί κατά τη μέτρηση της Ευκλείδειας απόστασης. Ας σημειωθεί ότι άλλα μέτρα απόστασης, όπως η απόσταση Mahalanobis (αναλύεται στο Κεφάλαιο 2.7.2) δεν παρουσιάζουν αυτό το μειονέκτημα, το οποίο παρουσιάζεται με την Ευκλείδεια απόσταση.



Σχήμα 2.9: Δισδιάστατα σημεία, όπου μία διάσταση έχει πολύ μεγαλύτερο εύρος τιμών από την άλλη.

Αν εξετάσουμε το εύρος τιμών μετά την τυποποίηση, η πρώτη διάσταση λαμβάνει τιμές στο διάστημα [-2.55, 2.46], ενώ η δεύτερη στο διάστημα [-2.20, 2.37]. Αν επιθυμούμε και οι δύο διαστάσεις να παίρνουν τιμές στο διάστημα [0,1], τότε για κάθε διάσταση ξεχωριστά εφαρμόζουμε το μετασχηματισμό:

$$X = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

□

2.4 Κβαντοποίηση Αριθμητικών Ιδιοτήτων

Κβαντοποίηση (quantization) καλείται ο μετασχηματισμός μίας συνεχούς ιδιότητας σε διακριτή. Στη βιβλιογραφία, εναλλακτικά του όρου κβαντοποίηση χρησιμοποιείται και ο όρος *διακριτοποίηση* (discretization), ενώ η ειδική περίπτωση του μετασχηματισμού συνεχούς ιδιότητας σε δυαδική καλείται *δυαδικοποίηση* (binarization).

Η ανάγκη για κβαντοποίηση προέρχεται από τον περιορισμό που θέτουν κάποιοι αλγόριθμοι εξόρυξης για επεξεργασία μόνο διακριτών ιδιοτήτων (π.χ., αλγόριθμοι εξόρυξης κανόνων συσχέτισης και ορισμένοι αλγόριθμοι κατηγοριοποίησης). Όμως εκτός αυτού η κβαντοποίηση είναι αρκετές φορές χρήσιμη, επειδή απλοποιεί την αναπαράσταση τόσο των δεδομένων εισόδου όσο και των αποτελεσμάτων.

Η κβαντοποίηση μίας συνεχούς ιδιότητας εφαρμόζεται απεικονίζοντας τις τιμές της σε ένα διακριτό πεδίο ορισμού. Ιδανική αντιστοίχηση είναι αυτή που επιτυγχάνει το βέλτιστο αποτέλεσμα για τον αλγόριθμο που στη συνέχεια θα εφαρμοσθεί επί της συνεχούς μεταβλητής. Για παράδειγμα, αν εφαρμοσθεί αλγόριθμος κατηγοριοποίησης, τότε το αποτέλεσμα θα μπορούσε να ελεγχθεί μέσω της ακρίβειας του αποτελέσματος. Όμως, τις περισσότερες φορές δεν είναι εφικτή η άμεση εφαρμογή τέτοιων κριτηρίων. Έτσι, προκύπτει η ανάγκη για αλγορίθμους κβαντοποίησης που θα στηρίζονται σε γενικότερα κριτήρια.

2.4.1 Αλγόριθμοι κβαντοποίησης

Μία γενική μέθοδος κβαντοποίησης περιλαμβάνει δύο στάδια:

i. τον ορισμό του διακριτού πεδίου ορισμού (πεπερασμένου ή αριθμήσιμου μη πεπερασμένου), και

ii. τον ορισμό της αντιστοίχησης από το συνεχές στο διακριτό πεδίο ορισμού.

Στη συνέχεια, θεωρούμε ότι το διακριτό πεδίο ορισμού είναι πεπερασμένο, ενώ είναι εύκολο να πραγματοποιηθεί η γενίκευση σε μη πεπερασμένα αριθμήσιμα πεδία. Ένα αλγοριθμικό πλαίσιο για την υλοποίηση των δύο σταδίων συνίσταται στα εξής βήματα:

1. Υπολόγισε το μέγεθος n του διακριτού πεδίου.
2. Ταξινόμησε τις τιμές της συνεχούς μεταβλητής.
3. Διαχώρισε τις τιμές αυτές σε n διακριτά διαστήματα, με καθορισμό $n-1$ σημείων διαχωρισμού.
4. Αντιστοίχησε ένα-προς-ένα κάθε διάστημα του συνεχούς πεδίου σε μία εκ των n τιμών του διακριτού πεδίου.

Τα Βήματα 2 και 4 είναι τετριμμένα. Επομένως, προκύπτει ότι για το Στάδιο i, πρέπει να ορίσουμε την τιμή του n (Βήμα 1), ενώ για το Στάδιο ii, πρέπει να ορίσουμε τον τρόπο διαχωρισμού των τιμών της συνεχούς μεταβλητής σε n διαστήματα (Βήμα 3). Στη συνέχεια παρουσιάζουμε δύο διαφορετικές θεωρήσεις για την υλοποίηση των Βημάτων 1 και 3. Η πρώτη ονομάζεται μη επιβλεπόμενη κβαντοποίηση, ενώ η δεύτερη ονομάζεται επιβλεπόμενη κβαντοποίηση.

Μη επιβλεπόμενη κβαντοποίηση

Η *μη επιβλεπόμενη κβαντοποίηση* (non supervised quantization) δεν χρησιμοποιεί άλλη πληροφορία πλην των τιμών της συνεχούς μεταβλητής. Στη μέθοδο αυτή, ο αριθμός των διαστημάτων n είναι είτε δεδομένος από το χρήστη, είτε ορίζεται μέσω ευρετικών μεθόδων (heuristics), όπως ο τύπος του Sturges:

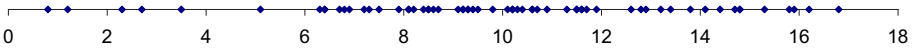
$$n \;=\; \log_2(M+1)$$

όπου M είναι το πλήθος των διαφορετικών τιμών των δεδομένων εισόδου.

Για το διαχωρισμό των τιμών της συνεχούς μεταβλητής, ορίζονται δύο διαφορετικοί τρόποι. Ο πρώτος τρόπος διαχωρισμού ονομάζεται *ισο-ευρύς* (equi-width), ενώ ο δεύτερος *ισο-υψής* (equi-heigth) ή *ισό-συχνος* (equi-frequency).

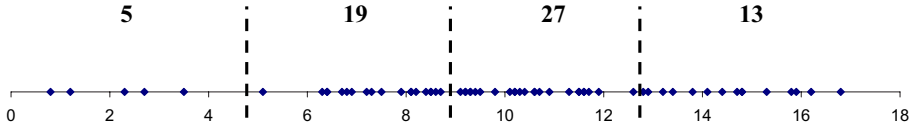
Ισο-ευρύς διαχωρισμός. Έστω X μία συνεχής ιδιότητα με M διαφορετικές τιμές. Αν $X_{\min}$ και $X_{\max}$ είναι αντιστοίχως η ελάχιστη και μέγιστη τιμή της X, τότε οι M διαφορετικές τιμές της X διαχωρίζονται σε διαστήματα ίσου εύρους, όπου το εύρος κάθε διαστήματος είναι σταθερό και ίσο με $W = (X_{\max} - X_{\min})/n$. Συνεπώς, τα προκύπτοντα σημεία διαχωρισμού είναι:

$X_{\min}+W, X_{\min}+2W, \ldots, X_{\min}+(n-1)W$. Για παράδειγμα, έστω ότι η ιδιότητα X περιλαμβάνει 64 τιμές στο διάστημα $[0.8, 16.8]$, οι οποίες απεικονίζονται στο Σχήμα 2.10.



Σχήμα 2.10: Συνεχής ιδιότητα με 64 τιμές.

Αν θέσουμε $n = 4$, τότε το εύρος W κάθε διαστήματος προκύπτει ίσο με 4. Το αποτέλεσμα απεικονίζεται στο Σχήμα 2.11 και περιλαμβάνει 4 διαστήματα: [0.8, 4,8), [4.8, 8.8), [8.8, 12.8), [12.8, 16.8]. Τα όρια κάθε διαστήματος απεικονίζονται με διακεκομμένη γραμμή. Επίσης επάνω από κάθε διάστημα αναγράφεται ο αριθμός των σημείων που εμπίπτουν σε αυτό. Παρατηρούμε ότι το εύρος των διαστημάτων είναι σταθερό αλλά οι αριθμοί των σημείων σε κάθε διάστημα είναι άνισοι. Μάλιστα, σε περιπτώσεις όπως αυτή του Σχήματος 2.11 προκύπτει ότι η κατανομή του αριθμού σημείων σε κάθε διάστημα είναι αρκετά ανομοιόμορφη (skewed). Αυτό μπορεί να αποτελέσει σημαντικό μειονέκτημα, καθώς δεν αντιπροσωπεύεται κάθε διάστημα από ισοδύναμο αριθμό σημείων. Επιπλέον, η παρουσία θορύβου μπορεί να επηρεάσει αρνητικά τα προκύπτοντα διαστήματα.



Σχήμα 2.11: Ισο-ευρής διαχωρισμός των 64 τιμών.

Τέλος, δύο επιπλέον ευρετικές μέθοδοι που εφαρμόζονται στον ισο-ευρύ διαχωρισμό για τον προσδιορισμό του n είναι ο τύπος των Friedman-Diaconis, ο οποίος ορίζεται ως:

$$n = \frac{X_{\max} - X_{\min}}{W}$$

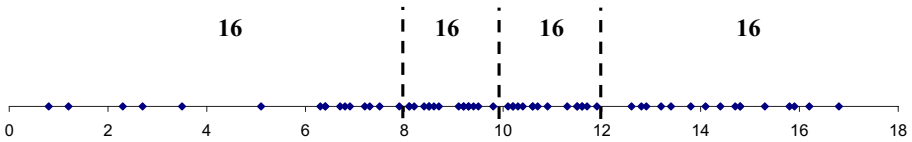
όπου $W = 2 * \text{IQR} * M^{-1/3}$ και $\text{IQR} = Q_3 - Q_1$, και ο τύπος του Scott, ο οποίος ορίζεται ως:

$$n = \frac{X_{\max} - X_{\min}}{W}$$

όπου $W = 3.5 \cdot s \cdot M^{-1/3}$, ενώ s είναι η τυπική απόκλιση.

Ισο-υψής διαχωρισμός. Μπορούμε να ακολουθήσουμε μία διαφορετική προσέγγιση για να αποφύγουμε το μειονέκτημα του ισο-ευρούς διαχωρισμού.

Έστω X μία συνεχής ιδιότητα με M διαφορετικές τιμές που διαχωρίζονται σε διαστήματα διαφορετικού εύρους, όπου το εύρος κάθε διαστήματος ορίζεται έτσι ώστε το πλήθος των σημείων που περιέχονται στα διαστήματα να είναι κατά το δυνατόν ίσα με M/n. Θεωρούμε και πάλι την ιδιότητα X με τις 64 τιμές που απεικονίζονται στο Σχήμα 2.10. Αν θέσουμε $n = 4$, τότε το αποτέλεσμα απεικονίζεται στο Σχήμα 2.12 και περιλαμβάνει 4 διαστήματα: [0.8, 7.9], [8.1, 9.8], [10.1, 11.9], [12.6, 16.8]. Παρατηρούμε ότι το εύρος των διαστημάτων δεν είναι πλέον σταθερό αλλά το πλήθος των σημείων σε κάθε διάστημα είναι πλέον ίσο. Έτσι, ο ισο-υψής διαχωρισμός προσαρμόζεται στην κατανομή των δεδομένων και δεν είναι τόσο ευαίσθητος στην παρουσία θορύβου. Για τους λόγους αυτούς, τις περισσότερες φορές ο ισο-υψής διαχωρισμός είναι προτιμότερος του ισο-ευρούς.



Σχήμα 2.12: Ισο-υψής διαχωρισμός των 64 τιμών.

Επιβλεπόμενη κβαντοποίηση

Σε κάποιες περιπτώσεις είναι δυνατόν κάθε τιμή της συνεχούς μεταβλητής να συνοδεύεται και από την πληροφορία της κλάσης όπου ανήκει. Σε αυτήν την περίπτωση, η μέθοδος της επιβλεπόμενης κβαντοποίησης μπορεί να εκμεταλλευθεί αυτήν την πληροφορία της κλάσης. Αυτό μπορεί να γίνει καθορίζοντας τα διαστήματα διαχωρισμού έτσι ώστε σε κάθε ένα από αυτά να προκύπτουν τιμές που, κατά το δυνατόν, ανήκουν στην ίδια κλάση. Ένας τρόπος ελέγχου της ομοιογένειας των τμημάτων σχετικά με τις κλάσεις που περιέχουν, είναι μέσω της ελαχιστοποίησης της εντροπίας.

Έστω M ο συνολικός αριθμός τιμών, ενώ κάθε τιμή ανήκει σε μία από k διακριτές κλάσεις. Θεωρώντας έναν (αυθαίρετο) διαχωρισμό σε n διαστήματα, έστω ότι στο διάστημα i εμπίπτουν m_i τιμές, ενώ $m_{i,j}$ είναι το πλήθος των τιμών στο διάστημα i, οι οποίες ανήκουν στην κλάση j. Ορίζουμε ως εντροπία e_i του διαστήματος i την εξής ποσότητα:

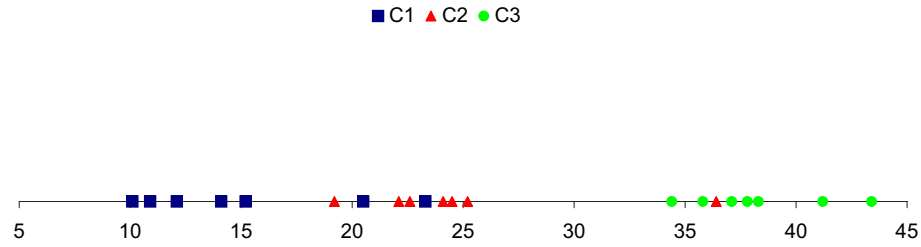
$$e_i \;=\; -\sum_{j=1}^{k} \frac{m_{i,j}}{m_i} \log_2 \frac{m_{i,j}}{m_i}. \tag{2.1}$$

Κατά τον υπολογισμό της Εξίσωσης 2.1, υποθέτουμε ότι $0 \log_2 0 = 0$. Όταν ένα διάστημα i περιέχει τιμές της ίδιας κλάσης, τότε προκύπτει ότι $e_i = 0$. Η μέγιστη τιμή του e_i προκύπτει όταν στο διάστημα i περιέχεται το ίδιος πλήθος τιμών από όλες τις κλάσεις. Η συνολική εντροπία e για όλα τα διαστήματα διαχωρισμού ορίζεται ως:

$$e = \sum_{i=1}^{n} \frac{m_i}{M} e_i. \tag{2.2}$$

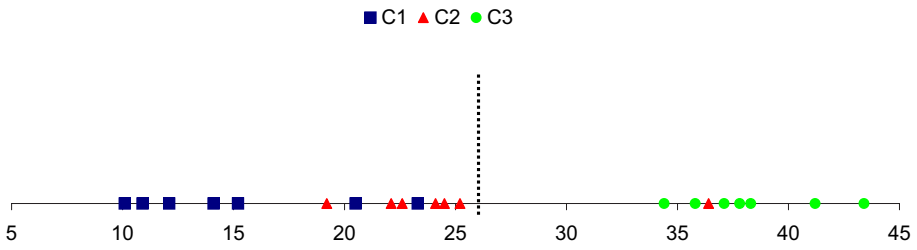
Με βάση τα προηγούμενα, αρχικά διαχωρίζουμε το διάστημα των τιμών σε δύο διαστήματα, έτσι ώστε να ελαχιστοποιείται η συνολική εντροπία e. Στη συνέχεια, για το διάστημα i με τη μέγιστη τιμή e_i εκ των δύο που προέκυψαν, εφαρμόζουμε αναδρομικά την ίδια διαδικασία. Τερματίζουμε τη διαδικασία είτε όταν φθάσουμε σε έναν προκαθορισμένο αριθμό διαστημάτων n, είτε όταν ικανοποιείται κάποιο άλλο κριτήριο.

Για παράδειγμα, το Σχήμα 2.13 περιλαμβάνει τις τιμές μίας συνεχούς ιδιότητας. Κάθε τιμή απεικονίζεται με διαφορετικό σχήμα (και χρώμα), αναλόγως της αντίστοιχης κλάσης. Η κλάση μίας τιμής μπορεί να είναι ίση με C_1, C_2 ή C_3. Υποθέτουμε ότι $n = 3$. Όπως φαίνεται στην εικόνα, υπάρχουν 3 ευδιάκριτες ομάδες τιμών. Η αριστερότερη είναι απολύτως ομογενής, η δεξιότερη περιέχει μόνο μία τιμή με διαφορετική κλάση από τις υπόλοιπες, ενώ η ενδιάμεση είναι η λιγότερο ομογενής. Διαισθητικά, ο διαχωρισμός θα θέλαμε να αντικατοπτρίζει αυτές τις ομάδες.



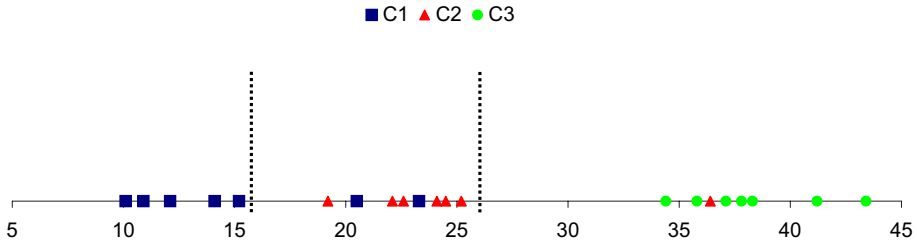
Σχήμα 2.13: Τιμές ιδιότητας και αντίστοιχες κλάσεις για κάθε τιμή.

Αν αρχικά διαχωρίσουμε σε δύο διαστήματα, η ελάχιστη συνολική εντροπία προκύπτει όταν το σημείο διαχωρισμού είναι η τιμή 25.2. Άρα, προκύπτουν δύο διαστήματα: [10.1, 25.2] και [34.4, 43.4], τα οποία απεικονίζονται στο Σχήμα 2.14. Σημειώνεται ότι, το σημείο διαχωρισμού μπορεί να τοποθετηθεί είτε στο άκρο του αριστερού διαστήματος (επί του σημείου διαχωρισμού), είτε στο μέσο μεταξύ των άκρων (δεξιού και αριστερού, αντιστοίχως) των δύο διαστημάτων.



Σχήμα 2.14: Διαχωρισμός σε 2 διαστήματα ελαχιστοποιώντας την εντροπία.

Στη συνέχεια, διαχωρίζουμε αναδρομικά το αριστερό διάστημα επειδή έχει τη μεγαλύτερη τιμή εντροπίας e_i. Το αποτέλεσμα απεικονίζεται στο Σχήμα 2.15. Η διαδικασία τερματίζεται, αφού θέλαμε διαχωρισμό σε 3 διαστήματα. Παρατηρώντας το αποτέλεσμα, προκύπτει ότι ο αλγόριθμος εντόπισε τις 3 ομάδες τιμών που διαισθητικά αντιληφθήκαμε ως καλύτερες.



Σχήμα 2.15: Αναδρομικός διαχωρισμός ελαχιστοποιώντας την εντροπία.

Τέλος, ας σημειωθεί ότι, για αυτόν τον αλγόριθμο είναι εφικτός ο αυτόματος υπολογισμός του n με βάση την αρχή της *Αρχής Περιγραφής Ελαχίστου Μήκους* (Minimum Description Length Principle - MDLP), η οποία θα αναλυθεί σε επόμενο κεφάλαιο.

2.5 Τυχαία Δειγματοληψία

Δεδομένου ενός συνόλου αντικειμένων, το οποίο καλείται πληθυσμός, *τυχαία δειγματοληψία* (random smapling) ονομάζεται η διαδικασία συλλογής ενός υποσυνόλου αντικειμένων του πληθυσμού, όπου κάθε αντικείμενο έχει μία γνωστή εκ των προτέρων πιθανότητα επιλογής. Συνήθως στη στατιστική, ο συνολικός πληθυσμός είτε είναι άγνωστος είτε είναι οικονομικά και χρονικά απαγορευτικό να μελετηθεί ολόκληρος. Επομένως, αντί του πληθυσμού, μελετάται ένα δείγμα

του, που παράγεται μέσω ενός δειγματοληπτικού πλαισίου (sampling frame).[3] Αντιθέτως, στην εξόρυξη δεδομένων ο πληθυσμός περιλαμβάνει ένα συγκεκριμένο σύνολο δεδομένων. Σε αυτήν την περίπτωση η τυχαία δειγματοληψία χρησιμοποιείται απλώς για τη δημιουργία ενός αντιπροσωπευτικού υποσυνόλου των δεδομένων, ώστε να μειωθεί ο χρόνος εκτέλεσης του αλγορίθμου εξόρυξης.

Η ανάπτυξη αλγορίθμων που κλιμακώνονται σε μεγάλους αριθμούς δεδομένων, είναι από τους βασικότερους στόχους στη γνωστική περιοχή της εξόρυξης δεδομένων. Επομένως, προκύπτει το ερώτημα αν είναι προτιμότερο να χρησιμοποιηθεί ένας κλιμακούμενος αλγόριθμος (π.χ., γραμμικής πολυπλοκότητας) σε ολόκληρο τον πληθυσμό, ή να χρησιμοποιηθεί ένας δαπανηρότερος (π.χ., τετραγωνικής πολυπλοκότητας) σε ένα δείγμα. Η τάση τα τελευταία χρόνια είναι προς την πρώτη κατεύθυνση. Εντούτοις, η τυχαία δειγματοληψία είναι χρήσιμη και έχει εφαρμοσθεί σε αρκετές περιπτώσεις. Επιπλέον, εκτός της μείωσης του μεγέθους των δεδομένων, η τυχαία δειγματοληψία έχει και το πλεονέκτημα της μείωσης του θορύβου, όπως θα αναλυθεί στη συνέχεια.

Λόγω του εύρους των εφαρμογών, έχουν αναπτυχθεί πάρα πολλές μέθοδοι τυχαίας δειγματοληψίας. Στη συνέχεια περιγράφουμε μόνο δύο από τις βασικότερες, την *απλή τυχαία δειγματοληψία* (simple random sampling) και τη *στρωματοποιημένη δειγματοληψία* (stratified sampling).

2.5.1 Απλή τυχαία δειγματοληψία

Με την απλή τυχαία δειγματοληψία, κάθε αντικείμενο έχει ίση πιθανότητα επιλογής σε σχέση με τα υπόλοιπα αντικείμενα του πληθυσμού. Επίσης, η επιλογή των αντικειμένων δεν ακολουθεί κάποια διάταξη που ενδεχομένως να υπάρχει στα δεδομένα. Για παράδειγμα, έστω τα ονόματα 100 φοιτητών, τα οποία αποθηκεύονται σε αλφαβητική διάταξη σε έναν πίνακα. Αν θέλουμε ένα απλό τυχαίο δείγμα 10 φοιτητών, τότε τυχαία (με ομοιόμορφη κατανομή) επιλέγουμε 10 αριθμούς μεταξύ 1 και 100, χωρίς κάποια διάταξη. Στη συνέχεια, από τον πίνακα ανακτούμε τα ονόματα που αντιστοιχούν στους 10 επιλεγμένους αριθμούς. Ως αποτέλεσμα, κάθε αρχικό όνομα έχει πιθανότητα ίση με $10/100 = 0.1$ για να

[3] Δειγματοληπτικό πλαίσιο είναι ένα μέσο ανάκτησης αντικειμένων του πληθυσμού. Π.χ., σε μία προεκλογική δημοσκόπηση, δειγματοληπτικό πλαίσιο μπορεί να αποτελέσει ο τηλεφωνικός κατάλογος. Αν το δειγματοληπτικό πλαίσιο δεν επιλεγεί σωστά, τότε το δείγμα δεν είναι αντιπροσωπευτικό. Για παράδειγμα, το 1936, μία προεκλογική δημοσκόπηση με δειγματοληπτικό πλαίσιο τη λίστα συνδρομητών ενός περιοδικού, είχε ως δείγμα ένα εκατομμύριο πολίτες των Η.Π.Α. Αν και το δείγμα ήταν πολύ μεγάλο, το δειγματοληπτικό πλαίσιο δεν ήταν σωστό, γιατί είχε σημαντικά περισσότερους ρεπουμπλικάνους. Έτσι, προβλέφθηκε λανθασμένα η νίκη των ρεπουμπλικάνων.

επιλεγεί στο δείγμα. Είναι ευνόητο ότι η διάταξη των ονομάτων δεν επηρέασε τη διαδικασία επιλογής.

Μία παραλλαγή της απλής τυχαίας δειγματοληψίας, είναι η *συστηματική (τυχαία) δειγματοληψία*. Στο προηγούμενο παράδειγμα, αρχικά επιλέγουμε το όνομα σε μία τυχαία γραμμή στον πίνακα, και στη συνέχεια επιλέγουμε το όνομα που βρίσκεται κάθε, π.χ., 10η γραμμή μετά, μέχρι να επιλέξουμε συνολικά 10 ονόματα (αν φθάσουμε στο τέλος του πίνακα, τότε συνεχίζουμε την αναζήτηση από την αρχή του πίνακα). Η συστηματική δειγματοληψία είναι χρονικά αποτελεσματικότερη σε σχέση με την απλή τυχαία δειγματοληψία. Όμως, ένα πιθανό μειονέκτημά της είναι ότι επηρεάζεται από τη διάταξη που υπάρχει στα δεδομένα, ειδικά στην περίπτωση που στα δεδομένα ενυπάρχει περιοδικότητα, οπότε και δεν θα προκύψει αντιπροσωπευτικό δείγμα.

Για την επιλογή ενός αντικείμενου κατά την απλή τυχαία δειγματοληψία, υπάρχουν δύο επιλογές:

- Το αντικείμενο δεν επιστρέφεται στον πληθυσμό, οπότε δεν μπορεί να επιλεγεί και πάλι. Αυτού του είδους η απλή τυχαία δειγματοληψία, ονομάζεται *χωρίς επανατοποθέτηση* (without replacement).

- Το αντικείμενο επιστρέφεται στον πληθυσμό, οπότε μπορεί και πάλι να επιλεγεί. Αυτού του είδους η απλή τυχαία δειγματοληψία, ονομάζεται *με επανατοποθέτηση* (with replacement).

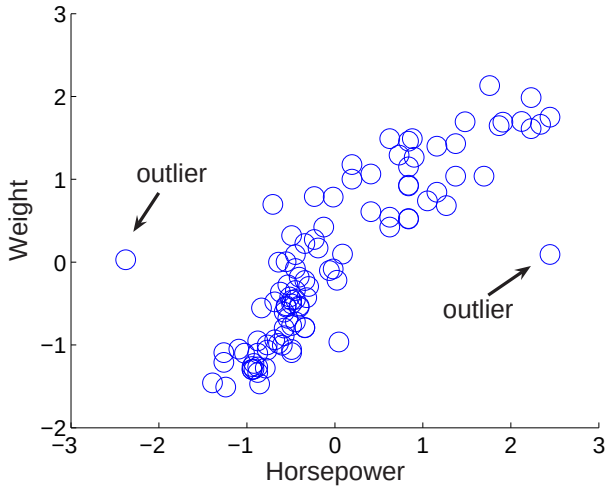
Στη δειγματοληψία με επανατοποθέτηση, όταν το μέγεθος του πληθυσμού είναι μεγάλο και το μέγεθος δείγματος μικρό, τότε η πιθανότητα επανεπιλογής του ιδίου αντικειμένου είναι μικρή. Σε αυτήν την περίπτωση, τα δείγματα που προκύπτουν με και χωρίς επανατοποθέτηση είναι σχεδόν όμοια. Η δειγματοληψία με επανατοποθέτηση έχει το πλεονέκτημα ότι αφήνει αμετάβλητη την πιθανότητα επιλογής κάθε αντικειμένου. Για αυτό το λόγο, χρησιμοποιείται περισσότερο όταν απαιτείται μαθηματική ανάλυση ιδιοτήτων του δείγματος. Αντιθέτως, αν η πιθανότητα επανεπιλογής αντικειμένων δεν είναι επιθυμητή, τότε χρησιμοποιείται η δειγματοληψία χωρίς επανατοποθέτηση.

Καθορισμός μεγέθους δείγματος

Μπορούμε να μετρήσουμε την ποιότητα ενός δείγματος σε σχέση με την ακρίβεια πρόβλεψης μίας ή και περισσοτέρων ιδιοτήτων του πληθυσμού. Για δεδομένη μέθοδο δειγματοληψίας, η ποιότητα του δείγματος εξαρτάται από το μέγεθός του. *Μέγεθος* (size) ενός δείγματος ονομάζουμε το πλήθος n των αντικειμένων του.

Υπάρχουν διάφορες ιδιότητες του πληθυσμού, μέσω των οποίων μπορούμε να ελέγξουμε την ποιότητα ενός δείγματος. Μία συχνά χρησιμοποιούμενη ιδιότητα είναι η διατήρηση στο δείγμα της μέσης τιμής που έχει μία τυχαία μεταβλητή στο συνολικό πληθυσμό. Μία άλλη ιδιότητα είναι η μεγιστοποίηση της ακρίβειας κατηγοριοποίησης που επιτυγχάνουμε με το δείγμα. Στα επόμενα παραδείγματα περιγράφεται η επιλογή μεγέθους δείγματος με βάση αυτές τις ιδιότητες.

Παράδειγμα 2.5.1 (Μέγεθος δείγματος με βάση τη μέση τιμή) Έστω ένα σύνολο δεδομένων με 100 τύπους αυτοκινήτων. Για κάθε τύπο έχουμε 2 ιδιότητες, την ιπποδύναμη (horsepower) και το βάρος (weight). Με κανονικοποίηση επιτυγχάνουμε η μέση τιμή και των 2 χαρακτηριστικών να είναι ίση με 0, ενώ η τυπική απόκλισή τους να είναι ίση με 1. Το αποτέλεσμα αποτυπώνεται γραφικά στο Σχήμα 2.16.



Σχήμα 2.16: 100 τύποι αυτοκινήτων με ιδιότητες την ιπποδύναμη και το βάρος.

Η ποιότητα του δείγματος θα ελεγχθεί με τη μέση τιμή της καθεμιάς από τις 2 ιδιότητες. Σε ένα απλό τυχαίο δείγμα, για μία ιδιότητα X, η μέση τιμή $\bar{X}$ στο δείγμα είναι μία τυχαία μεταβλητή που προσεγγιστικά ακολουθεί κανονική κατανομή. Μπορούμε να καθορίσουμε ένα διάστημα εμπιστοσύνης, θέτοντας ως a την πιθανότητα η μέση τιμή $\bar{X}$ στο δείγμα να απέχει κατά απόλυτη τιμή δ από τη μέση τιμή μ του πληθυσμού:

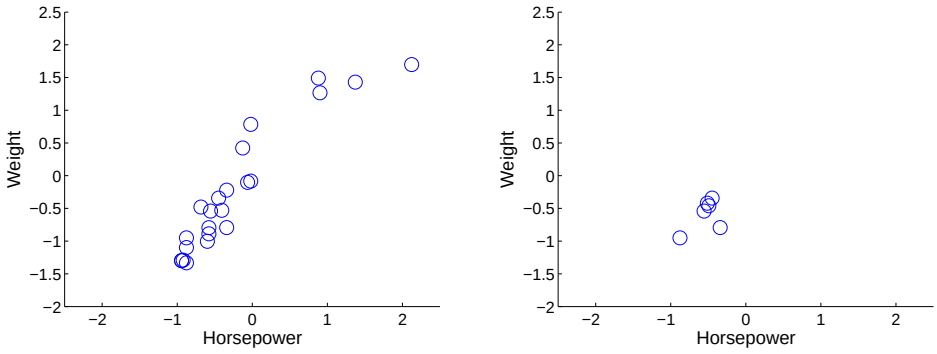
$$P(|\bar{X} - \mu| \geq \delta) \;=\; a$$

Σε αυτήν την περίπτωση, το μέγεθος δείγματος προκύπτει ίσο με:

$$n \approx \lceil z_a^2 \frac{\sigma^2}{\delta^2} \rceil \tag{2.3}$$

όπου σ είναι η τυπική απόκλιση της X στον πληθυσμό και z_a είναι κανονικοποιημένη τιμή της κανονικής κατανομής που αντιστοιχεί στην τιμή a.

Στο παράδειγμα, εξετάζουμε δύο τυχαίες μεταβλητές, την H που αντιστοιχεί στις τιμές της ιπποδύναμης, και την W που αντιστοιχεί στις τιμές του βάρους. Θα επιλέξουμε μέγεθος δείγματος με βάση τη μέση τιμή των δύο μεταβλητών. Επειδή στον πληθυσμό η τυπική απόκλιση σ είναι ίση με 1 και για τις δύο τυχαίες μεταβλητές, προκύπτει το ίδιο μέγεθος δείγματος και για τις δύο. Οπότε, στη συνέχεια, εξετάζουμε την H. Επιλέγουμε $a = 95\%$ (προκύπτει ότι $z_a = 2$), ενώ θέτουμε δύο τιμές για το δ. Η πρώτη είναι ίση με $\delta = 0.3$, οπότε από την Εξίσωση 2.3 προκύπτει $n = 23$, ενώ η δεύτερη είναι ίση με $\delta = 0.6$, οπότε προκύπτει $n = 6$. Τα δείγματα που παράγονται στις δύο περιπτώσεις, απεικονίζονται στα Σχήματα 2.17α και 2.17β, αντιστοίχως.



(α) (β)

Σχήμα 2.17: Παραγόμενα δείγματα για (α) $n = 23$, (β) $n = 6$.

Από το δείγμα του Σχήματος 2.17α προκύπτει ότι $\bar{H} = -0.21$ και $\bar{W} = -0.26$. Δεδομένου ότι στον πληθυσμό ισχύει: $\mu_H = \mu_W = 0$, και οι δύο τιμές είναι εντός του απαιτούμενου εύρους $\delta = 0.3$. Στο δείγμα του Σχήματος 2.17β προκύπτει ότι $\bar{H} = -0.54$ και $\bar{W} = -0.58$. Προφανώς, και οι δύο τιμές είναι επίσης εντός του απαιτούμενου εύρους $\delta = 0.6$.

Και στα δύο δείγματα, η μέση τιμή διατηρείται εντός του απαιτούμενου εύρους δ. Όμως αυτό το χαρακτηριστικό δεν είναι αρκετό για να καθορίσει την

ποιότητα ενός δείγματος, επειδή εξαρτάται από το αν το απαιτούμενο εύρος είναι ικανό να παράγει ένα αντιπροσωπευτικό δείγμα. Στο δεύτερο δείγμα, το απαιτούμενο εύρος είναι διπλάσιο από ότι στο πρώτο. Επομένως, συγκρίνοντας τα δύο δείγματα, το δεύτερο είναι περισσότερο αντιπροσωπευτικό από το πρώτο. Στον πληθυσμό, οι τιμές των H και P παρουσιάζουν μία θετική συσχέτιση που εκφράζεται από την τιμή της συνδιασποράς (covariance) τους $cov(H, P) = E[XY] = 0.85$. Η θετική συσχέτιση διατηρείται στο δείγμα του Σχήματος 2.17α, καθώς η συνδιασπορά είναι ίση με 0.74. Αντιθέτως, στο δείγμα του Σχήματος 2.17β η συσχέτιση δεν είναι θετική, επειδή η συνδιασπορά είναι ίση με 0.02. Άρα, το δεύτερο δείγμα δεν αποτυπώνει τη θετική συσχέτιση μεταξύ των H και P.

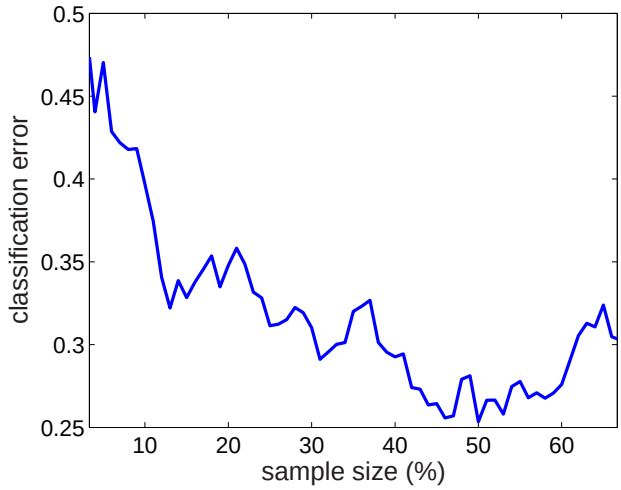
Ως συμπέρασμα, αν θέσουμε επαρκές διάστημα εμπιστοσύνης (π.χ., $a = 95\%$) και εύρος (π.χ., $\delta = 0.3$), μπορεί να παραχθεί ένα αντιπροσωπευτικό δείγμα. Φυσικά, μεγαλύτερα διαστήματα εμπιστοσύνης και μικρότερα εύρη, οδηγούν σε αντιπροσωπευτικότερα δείγματα. Όμως, τα δείγματα αυτά θα είναι μεγαλύτερα, ακυρώνοντας έτσι το πλεονέκτημα της μείωσης του χρόνου εκτέλεσης ενός αλγορίθμου. Επομένως, η επιλογή μεγέθους δείγματος πρέπει να οδηγεί σε όσο το δυνατόν μικρότερα και όσο το δυνατόν αντιπροσωπευτικότερα δείγματα, όπου όμως τα δύο αυτά κριτήρια είναι αντικρουόμενα.

Τέλος, στο δείγμα του Σχήματος 2.17α μπορούμε να παρατηρήσουμε τη μείωση του θορύβου. Στον πληθυσμό του Σχήματος 2.16 υπάρχουν μερικά σημεία που δεν ακολουθούν την κατανομή των υπολοίπων και δηλώνονται ως outliers. Εξαιτίας της δειγματοληψίας, υπάρχει η πιθανότητα κάποια από αυτά τα σημεία outliers να μην επιλεγούν στο δείγμα. Πράγματι, αυτά τα σημεία δεν εμφανίζονται στο Σχήμα 2.17α. Επομένως, όπως προαναφέρθηκε, η δειγματοληψία παρουσιάζει το πλεονέκτημα της μείωσης του θορύβου. □

Παράδειγμα 2.5.2 (Μέγεθος δείγματος με βάση την ακρίβεια) Αν ο πληθυσμός περιέχει αντικείμενα που ανήκουν σε κλάσεις, τότε μπορούμε να ελέγξουμε την ποιότητα του δείγματος με βάση την ακρίβεια που επιτυγχάνουμε όταν εκπαιδεύουμε έναν κατηγοριοποιητή με το δείγμα αυτό (η κατηγοριοποίηση θα αναλυθεί στο Κεφάλαιο 6).

Ως παράδειγμα, χρησιμοποιούμε το σύνολο δεδομένων Iris και λαμβάνουμε δείγματα διαφόρων μεγεθών. Για κάθε δείγμα, εκπαιδεύουμε έναν κατηγοριοποιητή που κατατάσσει λουλούδια από το σύνολο Iris σε ένα από τα 3 είδη. Μετρούμε την πιθανότητα λανθασμένης κατάταξης σε σχέση με το μέγεθος του δείγματος εκπαίδευσης του κατηγοριοποιητή. Τα αποτελέσματα απεικονίζονται στο Σχήμα 2.18.

Παρατηρούμε ότι με δείγματα μικρού μεγέθους προκύπτει μεγάλο σφάλμα



Σχήμα 2.18: Πιθανότητα λανθασμένης κατάταξης ως προς μέγεθος δείγματος.

κατηγοριοποίησης. Αντιθέτως, καθώς το μέγεθος του δείγματος μεγαλώνει, τόσο το σφάλμα μειώνεται. Αυτό συνεχίζεται μέχρι ενός σημείου, οπότε με μεγαλύτερα μεγέθη δείγματος το σφάλμα και πάλι αυξάνεται. Η τελευταία παρατήρηση εξηγείται από το γεγονός ότι ο κατηγοριοποιητής εκπαιδεύεται υπερβολικά για το συγκεκριμένο δείγμα και δεν μπορεί να αντιμετωπίσει καλά αντικείμενα που δεν ανήκουν στο δείγμα. Το φαινόμενο αυτό ονομάζεται overfiting και θα αναλυθεί σε επόμενο κεφάλαιο. Αυτό που έχει σημασία είναι να παρατηρήσουμε ότι αναλόγως του σφάλματος που μπορούμε να ανεχθούμε, αντιστοίχως θα επιλέξουμε και το μέγεθος του δείγματος. □

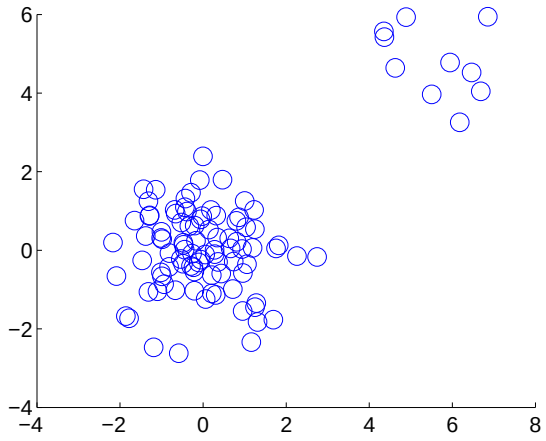
2.5.2 Στρωματοποιημένη δειγματοληψία

Συχνά ο πληθυσμός μπορεί να περιέχει αντικείμενα που χωρίζονται σε ομάδες κατά ένα φυσικό τρόπο. Κάθε αντικείμενο ανήκει σε μία και μόνη ομάδα που ονομάζεται *στρώμα* (stratum), ενώ ο διαχωρισμός των αντικειμένων στις ομάδες ονομάζεται *στρωματοποίηση* (stratification).

Αν το πλήθος των αντικειμένων διαφέρει σημαντικά μεταξύ των ομάδων, τότε η απλή τυχαία δειγματοληψία μπορεί να παράγει δείγμα όπου δεν αντιπροσωπεύονται ικανοποιητικά όλες οι ομάδες. Για να αντιμετωπισθεί το πρόβλημα αυτό, χρησιμοποιείται η στρωματοποιημένη (τυχαία) δειγματοληψία, όπου εφαρμόζουμε απλή τυχαία δειγματοληψία σε κάθε ομάδα ξεχωριστά. Το μέγεθος

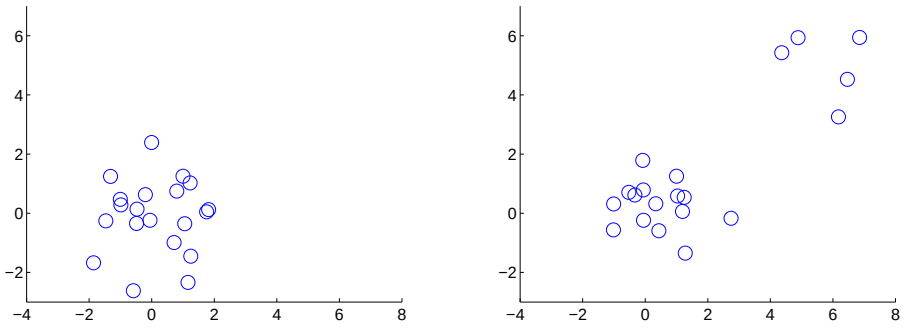
δείγματος κάθε ομάδας μπορεί να καθορισθεί με διάφορα κριτήρια. Στην απλή περίπτωση, το μέγεθος δείγματος από κάθε ομάδα, είναι ανάλογο του μεγέθους της ομάδας. Π.χ., αν έχουμε πληθυσμό 100 ατόμων, όπου τα υγιή είναι 90% και τα ασθενή 10%, και επιθυμούμε ένα δείγμα 10 ατόμων, τότε θα έπρεπε να επιλέξουμε $0.9 \cdot 10/100 = 9$ άτομα από την πρώτη ομάδα και $0.1 \cdot 10/100 = 1$ άτομο από τη δεύτερη ομάδα. Σε μία ακραία περίπτωση, θα μπορούσαμε να λάβουμε το ίδιο μέγεθος δείγματος για όλες τις ομάδες. Με τη λογική αυτή, στο προηγούμενο παράδειγμα θα επιλέγαμε 5 υγιή και 5 ασθενή άτομα. Ένα άλλο κριτήριο είναι να λάβουμε μεγαλύτερα δείγματα από τις ομάδες με μεγαλύτερη διασπορά. Το πλεονέκτημα της στρωματοποιημένης δειγματοληψίας περιγράφεται στο επόμενο παράδειγμα.

Παράδειγμα 2.5.3 (Στρωματοποιημένη δειγματοληψία) Ας υποθέσουμε ότι έχουμε εντοπίσει 100 τοποθεσίες φωλιών πουλιών, που ανήκουν σε δύο σπάνια είδη. Το πρώτο είδος περιλαμβάνει 90 πουλιά, ενώ το δεύτερο 10. Οι φωλιές προσδιορίζονται ως σημεία σε ένα δισδιάστατο σύστημα συντεταγμένων. Στον πληθυσμό, οι θέσεις των φωλιών του πρώτου είδους ακολουθούν κανονική κατανομή γύρω από το σημείο (0,0), ενώ του δεύτερου είδους ακολουθούν κανονική κατανομή γύρω από το σημείο (5,5). Οι συντεταγμένες των 100 φωλιών απεικονίζονται στο Σχήμα 2.19. Μπορούμε να αναζητούμε τη θέση κάθε φωλιάς ξεχωριστά μέσω κάποιου συστήματος ανίχνευσης. Οι θέσεις των φωλιών αρχικά μας είναι άγνωστες.



Σχήμα 2.19: Σύνολο 100 σημείων σε 2 ομάδες.

Θέλουμε να μελετήσουμε τις θέσεις των 100 φωλιών. Επειδή δεν επαρκεί ο χρόνος να αναζητήσουμε και να βρούμε και τις 100 φωλιές, αποφασίζουμε να σχηματίσουμε ένα δείγμα 20%, δηλαδή 20 φωλιών που θα εξετασθούν κατά την έρευνα. Αρχικά, σχηματίζουμε ένα απλό τυχαίο δείγμα 20 πουλιών και από τα δύο είδη. Στο Σχήμα 2.20α απεικονίζονται οι θέσεις των φωλιών των επιλεγμένων πουλιών. Επιλέχθηκαν τυχαία 20 πουλιά από το πρώτο είδος και κανένα από το δεύτερο. Επομένως, το δεύτερο είδος δεν αντιπροσωπεύεται στο δείγμα.



(α) (β)

Σχήμα 2.20: Παραγόμενα δείγματα με: (α) απλή τυχαία δειγματοληψία, (β) στρωματοποιημένη δειγματοληψία.

Στη συνέχεια εφαρμόζουμε στρωματοποιημένη δειγματοληψία. Έστω ότι λόγω προγενέστερης γνώσης, αναμένουμε ότι η διασπορά στις φωλιές του δεύτερου είδους είναι μεγαλύτερη από ότι του πρώτου. Λαμβάνουμε δείγμα μη αναλογικό ως προς τα μεγέθη των δύο ειδών, ώστε να αποτυπωθεί η διαφορά στις διασπορές. Επομένως, εφαρμόζουμε ξεχωριστά απλή τυχαία δειγματοληψία στη πρώτη ομάδα και επιλέγουμε 15 πουλιά, και ξεχωριστά στη δεύτερη επιλέγοντας 5 πουλιά. Οι θέσεις του δείγματος απεικονίζεται στο Σχήμα 2.20β. Αντιθέτως από την απλή τυχαία δειγματοληψία, η στρωματοποιημένη δειγματοληψία δημιούργησε δείγμα όπου και τα δύο είδη αντιπροσωπεύονται ικανοποιητικά. □

2.6 Μείωση Αριθμού Διαστάσεων

Συχνά το σύνολο δεδομένων μπορεί να έχει μεγάλο αριθμό διαστάσεων. Για παράδειγμα, τα σύνολα δεδομένων σε πολυμεσικές εφαρμογές έχουν αρκετές