Preprocesarea datelor

Preprocesarea datelor

- Curatarea datelor
  - Date lipsa, date incorecte, valori extreme (outliers)
- Integrarea datelor
- Transformarea datelor
- Reducerea datelor

Curatarea datelor

Customer ID	Zip	Gender	Income	Age	Marital Status	Transaction Amount
1001	10048	M	75000	C	M	5000
1002	J2S7K7	F	−40000	40	W	4000
1003	90210		10000000	45	S	7000
1004	6269	M	50000	0	S	1000
1005	55101	F	99999	30	D	3000

Probabil un cod postal al unei regiuni ce nu apartine SUA

Probabil codul postal 06269

Valoare lipsa

Eroare

Ar putea fi un cod pentru anomalii

Outlier (valoare extrema)

Valoare categoriala

Este clara semnificatia acestor simboluri?

Suntem siguri ca unitatea de masura este clara?

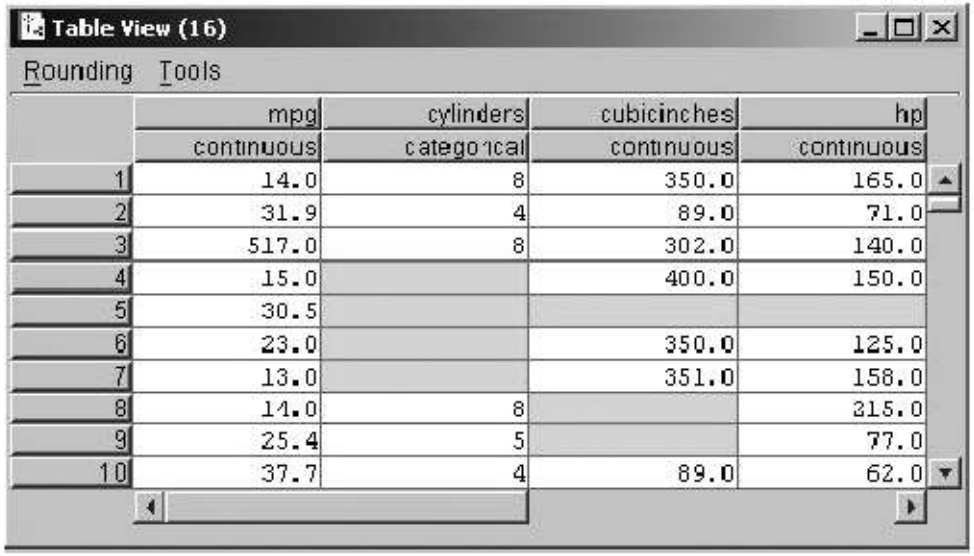
Curatarea datelor

- Motive pentru date lipsa, date inconsistente
  - discutii

Curatarea datelor
Date lipsa

Setul de date CARS (www.sgi.com/tech/mlc/db)

- Informatii despre 261 automobile fabricate intre 1970 si 1980
- Software-ul folosit pentru analiza valorilor lipsa este Insightful Miner (Insightful Corporation, **www.insightful.com**)



	mpg	cylinders	cubicinches	hp
	continuous	categorical	continuous	continuous
1	14.0	8	350.0	165.0
2	31.9	4	89.0	71.0
3	517.0	8	302.0	140.0
4	15.0		400.0	150.0
5	30.5			
6	23.0		350.0	125.0
7	13.0		351.0	158.0
8	14.0	8		315.0
9	25.4	5		77.0
10	37.7	4	89.0	62.0

Figure 2.1 Some of our field values are missing!

Referinta figura: D. Larose, *Discovering Knowledge in Data. An Introduction to Data Mining*, John Wiley & Sons, 2005.

Curatarea datelor
Date lipsa

- Omiterea din analiza a inregistrarilor sau a campurilor cu valori lipsa
  - Asta ar duce la o favorizare a unui subset de date
  - E o mare pierdere sa ignori informatia din alte campuri ale unei inregistrari doar pentru ca unele campuri au valori lipsa

- Inlocuirea cu o valoarea determinata prin diferite metode:
  - O constanta, specificata de analist
  - Media valorilor acelui camp (pentru variabile numerice) sau modul (pentru variabile categoriale)
  - O valoare generata aleator

- Inlocuirea cu cea mai potrivita valoare, tinand cont de valorile celorlalte atribute ale inregistrarii

Utilizatorii finali ai rezultatelor trebuie sa cunoasca ce fel de metoda a fost aplicata pentru a interpreta corect rezultatele!!!

Curatarea datelor
Date lipsa

Table View (17)

Rounding Tools

	mpg	cylinders	cubicinches	hp
	continuous	categorical	continuous	continuous
1	14.00	8	350.00	165.00
2	31.90	4	89.00	71.00
3	517.00	8	302.00	140.00
4	15.00	Missing	400.00	150.00
5	30.50	Missing	0.00	0.00
6	23.00	Missing	350.00	125.00
7	13.00	Missing	351.00	158.00
8	14.00	8	0.00	215.00
9	25.40	5	0.00	77.00
10	37.70	4	89.00	62.00

Figure 2.2 Replacing missing field values with user-defined constants.

POSIBILA PROBLEMA: programul de data mining ar putea interpreta in mod gresit ca aceste valori formeaza un concept interesant

Referinta figura: D. Larose, *Discovering Knowledge in Data. An Introduction to Data Mining*, John Wiley & Sons, 2005.

Curatarea datelor
Date lipsa

Table View (17)

Rounding Tools

	mpg	cylinders	cubicinches	hp
	continuous	categorical	continuous	continuous
1	14.00	8	350.00	165.00
2	31.90	4	89.00	71.00
3	517.00	8	302.00	140.00
4	15.00	4	400.00	150.00
5	30.50	4	200.65	106.53
6	23.00	4	350.00	125.00
7	13.00	4	351.00	158.00
8	14.00	8	200.65	215.00
9	25.40	5	200.65	77.00
10	37.70	4	89.00	62.00

Figure 2.3 Replacing missing field values with means or modes.

POSIBILA PROBLEMA:
Daca exista multe valori lipsa si acestea sunt inlocuite cu media, atunci dispersia va fi redusa artificial

Referinta figura: D. Larose, *Discovering Knowledge in Data. An Introduction to Data Mining*, John Wiley & Sons, 2005.

Curatarea datelor
Date lipsa

Table View (17)

Rounding Tools

	mpg	cylinders	cubicinches	hp
	continuous	categorical	continuous	continuous
1	14.00	8	350.00	165.00
2	31.90	4	89.00	71.00
3	517.00	8	302.00	140.00
4	15.00	8	400.00	150.00
5	30.50	4	144.15	116.55
6	23.00	4	350.00	125.00
7	13.00	6	351.00	158.00
8	14.00	8	323.45	215.00
9	25.40	5	81.84	77.00
10	37.70	4	89.00	62.00

Figure 2.4 Replacing missing field values with random draws from the distribution of the variable.

Referinta figura: D. Larose, *Discovering Knowledge in Data. An Introduction to Data Mining*, John Wiley & Sons, 2005.

Curatarea datelor
Identificarea clasificarilor gresite

- the frequency distribution of the categorical variable *origin*

Origin	Count
USA	1
France	1
US	156
Europe	46
Japan	51

Doua inregistrari au fost clasificate gresit in ceea ce priveste tara de fabricatie. USA ar trebui inlocuit cu US, iar France ar trebui inlocuit cu Europe

Referinta figura: D. Larose, *Discovering Knowledge in Data. An Introduction to Data Mining*, John Wiley & Sons, 2005.

Curatarea datelor

Identificarea valorilor extreme (outliers)

- *Outliers* = valori care se afla la extremitatile intervalelor din care fac parte valorile posibile ale unui atribut

- Identificarea lor este importanta deoarece:
  - Ar putea reprezenta erori
  - Chiar daca este o valoare valida si nu o eroare, anumite metode statistice sunt sensibile la prezenta valorilor extreme si ar putea da rezultate incorecte

Curatarea datelor

Identificarea valorilor extreme (outliers)

HISTOGRAMA

- Setul de date contine un automobil care are o greutate de 192.5 pounds
- vom tinde sa ne indoim de validitatea acestei informatii
- toate celelalte automobile au pentru greutate valori intregi, fara zecimale
- putem presupune ca greutatea a fost de fapt 1925 pounds, virgula fiind inserata din greseala
- nu putem fi totusi absolut siguri, de aceea sunt necesare investigatii suplimentare

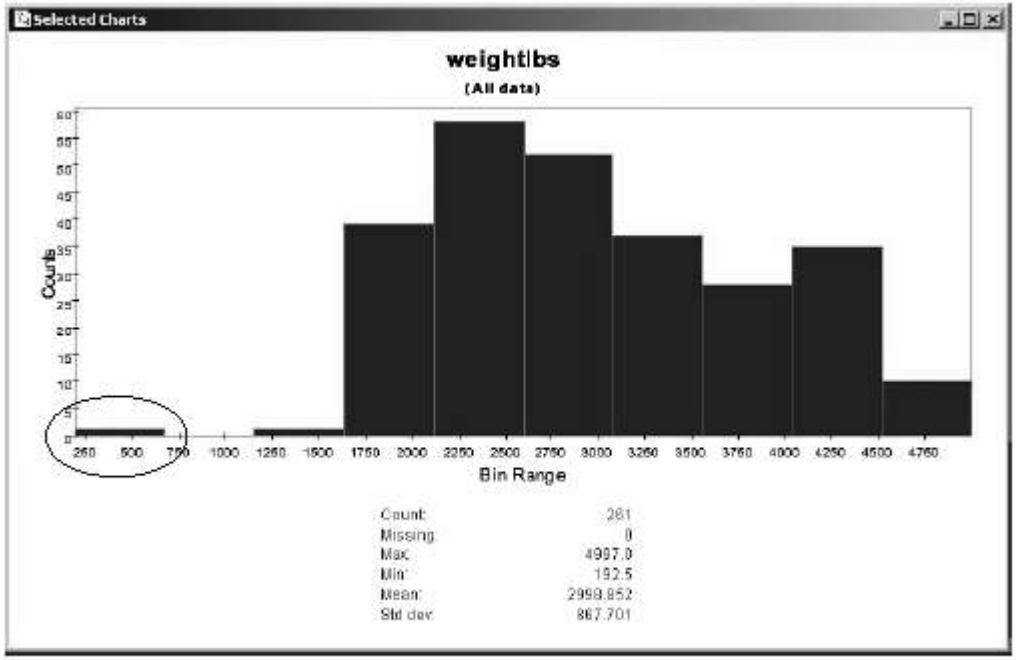


Figure 2.5 Histogram of vehicle weights: can you find the outlier?

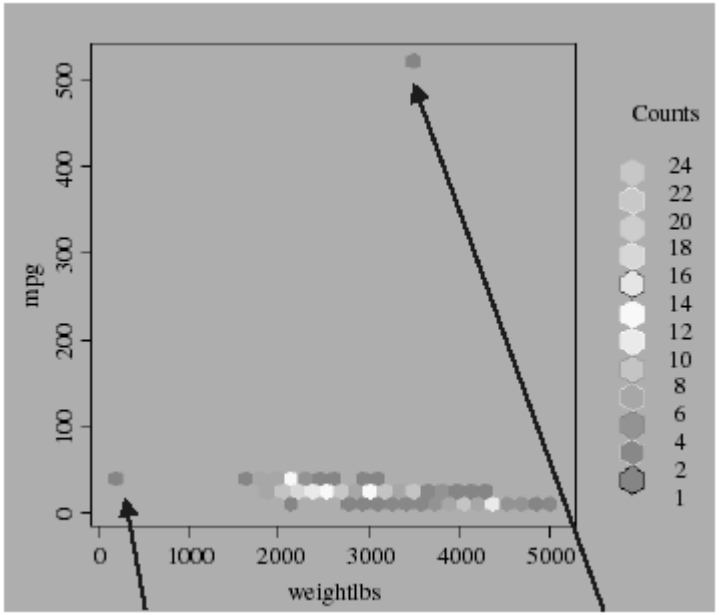
Referinta figura: D. Larose, *Discovering Knowledge in Data. An Introduction to Data Mining*, John Wiley & Sons, 2005.

Curatarea datelor

Identificarea valorilor extreme (outliers)

DIAGRAMA DE IMPRASTIERE BIDIMENSIONALA

- Pot ajuta la identificarea valorilor extreme pentru mai mult de o variabila
- se pot observa foarte usor cele 2 valori extreme
- cea din stanga este automobilul cu greutatea de numai 192.5 pounds
- in partea de sus este un automobil care consuma 1 galon/500 mile



Referinta figura: D. Larose, *Discovering Knowledge in Data. An Introduction to Data Mining*, John Wiley & Sons, 2005.

Integrarea datelor

- Combinarea datelor din mai multe surse intr-un depozit de date coerent, cum ar fi data warehouse.

Integrarea datelor

- *Structura bazelor de date*
  - Unele atribute pot sa aiba nume diferite in baze de date diferite (exemplu: *customer_id* intr-o baza de date, si *cust_id* in alta)
  - Codificarea unor date pentru campul *pay type* pot fi *H* si *S intr-o baza de date,* si *1* si *2* in alta).
  - Acelasi nume poate fi inregistrat ca “Bill” intr-o baza de date, “William” in alta, si “B.” intr-o a treia
- *Redundanta*
  - Un atribut poate fi redundant daca el poate fi derivat din alte atribute sau seturi de atribute (daca exista multe date redundante, ar putea sa incetineasca sau sa dea rezultate confuze in procesul de data mining)
- *Duplicare*
- *Detectarea si rezolvarea conflictelor de date*

Integrarea datelor
Detectarea redundantelor

- Analiza de corelatie
  - Fiind date 2 atribute, o astfel de analiza va masura cat de tare un atribut il implica pe celalalt, bazandu-ne pe datele existente
  - Pentru atribute numerice putem evalua corelatia dintre 2 atribute *A* si *B*, calculand **coeficientul de corelatie** (cunoscut si ca si *Pearson's product moment coefficient)*
  - Pentru valori discrete (categoriale) o corelatie intre doua atribute poate fi detectata prin testul **chi-square.**

Integrarea datelor
Detectarea redundantelor

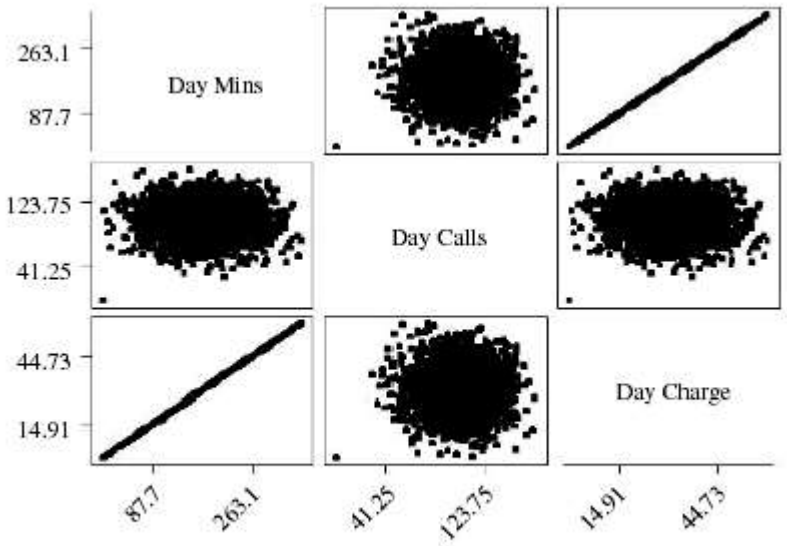


Figure 3.2 Matrix plot of *day minutes*, *day calls*, and *day charge*.

Aceasta diagrama este realizata folosind pachetul statistic **Minitab**

- Nu pare sa existe vreo relatie intre *day minutes* si *day calls* sau intre *day calls* si *day charge*.

- Pe de alta parte, exista o relatie perfect liniara intre *day minutes* si *day charge*, indicand ca *day charge* este o functie simpla liniara doar de *day minutes*.

Deoarece *day charge* este perfect corelata cu *day minutes*, ar trebui sa eliminam una din cele 2 variabile.

Referinta figura: D. Larose, *Discovering Knowledge in Data. An Introduction to Data Mining*, John Wiley & Sons, 2005.

Transformarea datelor

- Datele sunt transformate in forme potrivite pentru procesul de data mining

- Transformarea datelor poate sa implice urmatoarele:
  - *Smoothing*
  - *Agregare*
  - *Generalizare*
  - *Normalizare,* unde valorile atributelor sunt scalate astfel incat toate sa apartina unui anumit interval cum ar fi [-1.0, 1.0], sau [0.0, 1.0].
  - *Construirea de atribute*, unde noi atribute sunt construite sau adaugate pentru a ajuta in procesul de data mining

Transformarea datelor

Normalizare

- De ce este necesara normalizarea variabilelor numerice?
- Domeniile variabilele variaza foarte mult de la o variabila la alta
- Pentru unii algoritmi de data mining, aceste diferente vor duce la o tendinta ca variabila cu domeniu mai mare sa aiba o anumita influenta nejustificata asupra rezultatelor
- Metode:
  - Normalizarea Min–Max
  - Standardizarea *Z*-Score

Transformarea datelor

Normalizarea Min–Max

$$X^* = \frac{X - \min(X)}{\text{range}(X)} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

- In urma normalizarii min–max valorile vor fi intre 0 si 1

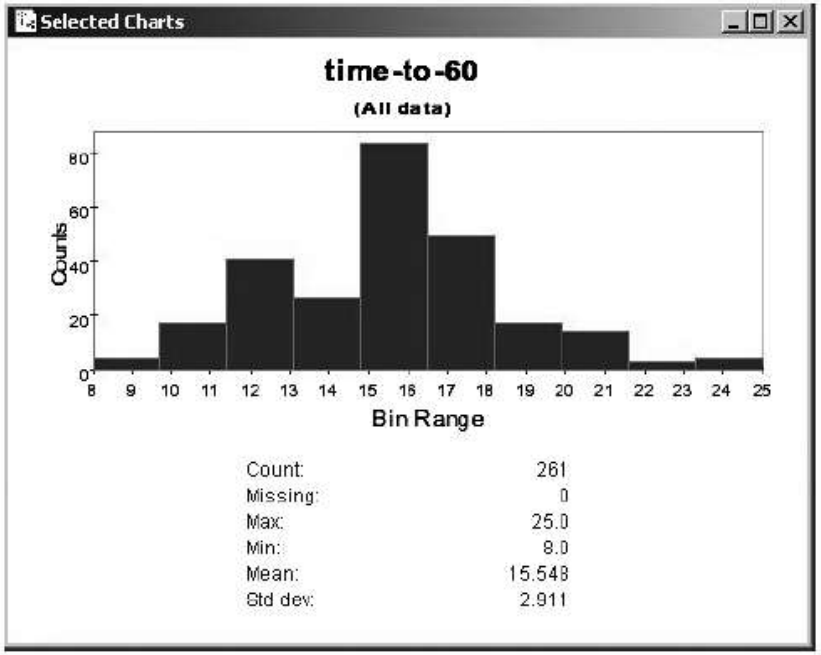


Figure 2.7 Histogram of *time-to-60*, with summary statistics.

Referinta figura: D. Larose, *Discovering Knowledge in Data. An Introduction to Data Mining*, John Wiley & Sons, 2005.

Transformarea datelor

Standardizare *Z*-Score

$$X^* = \frac{X - \text{mean}(X)}{\text{SD}(X)}$$

- In urma standardizarii *Z*-score valorile vor fi de obicei intre –4 si 4, media avand dupa standardizare valoarea 0
- putem sa detectam valori extreme (outliers) folosind standardizarea ***Z*-score**

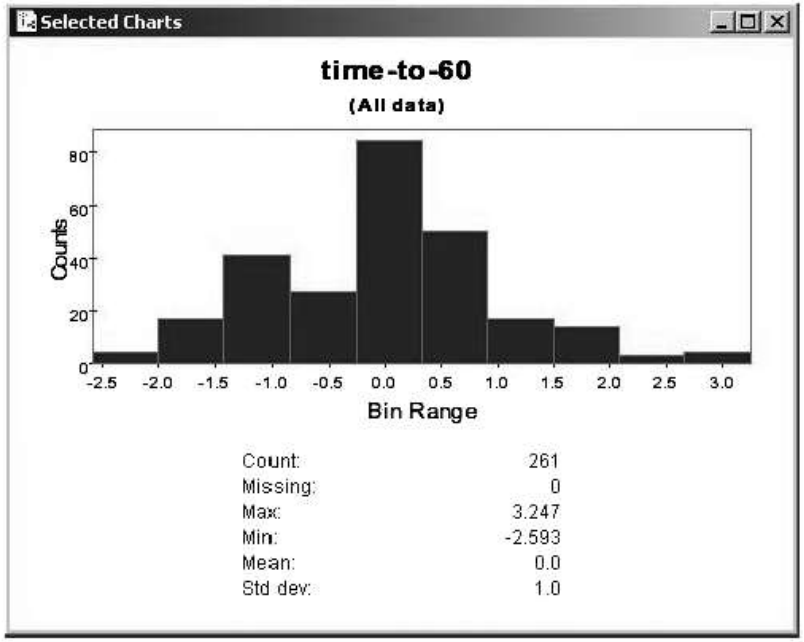


Figure 2.8 Histogram of *time-to-60* after *Z*-score standardization.

Referinta figura: D. Larose, *Discovering Knowledge in Data. An Introduction to Data Mining*, John Wiley & Sons, 2005.

Transformarea datelor

Construirea de noi atribute

- Noi atribute sunt create din atributele existente si adaugate pentru a ajuta la imbunatatirea acuratetei si intelegerii structurii datelor
- De exemplu, am putea sa adaugam atributul *area* bazat pe atributele *height* si *width*
- Prin combinarea atributelor se pot descoperi informatii lipsa despre relatiile intre atribute, si asta poate ajuta in procesul de data mining

Reducerea datelor

• Selectarea unor subseturi de date interesante pentru investigatii ulterioare (vezi figura alaturata)
  - clientii cu multe *day minutes si multe evening minutes* sunt mai predispusi la schimbarea furnizorului de telefonie mobila

• Selectarea doar a unor anumite seturi de atribute

• *Binning* = discretizarea variabilelor numerice intr-un set de clase care sunt potrivite pentru analiza
  - de ex, numarul *day minutes* poate fi categorizat (binned) in 3 clase: *low*, *medium*, and *high*.

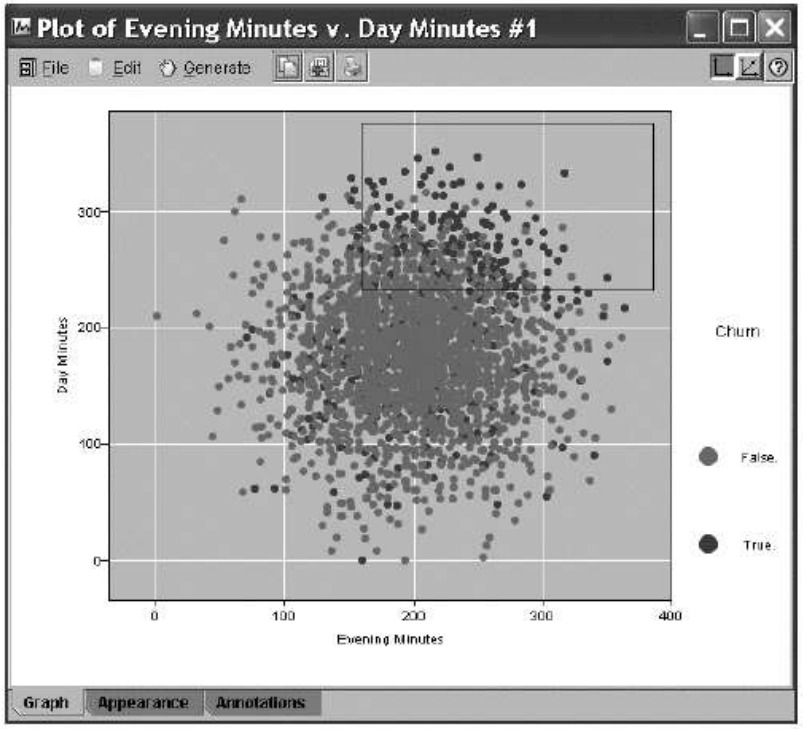
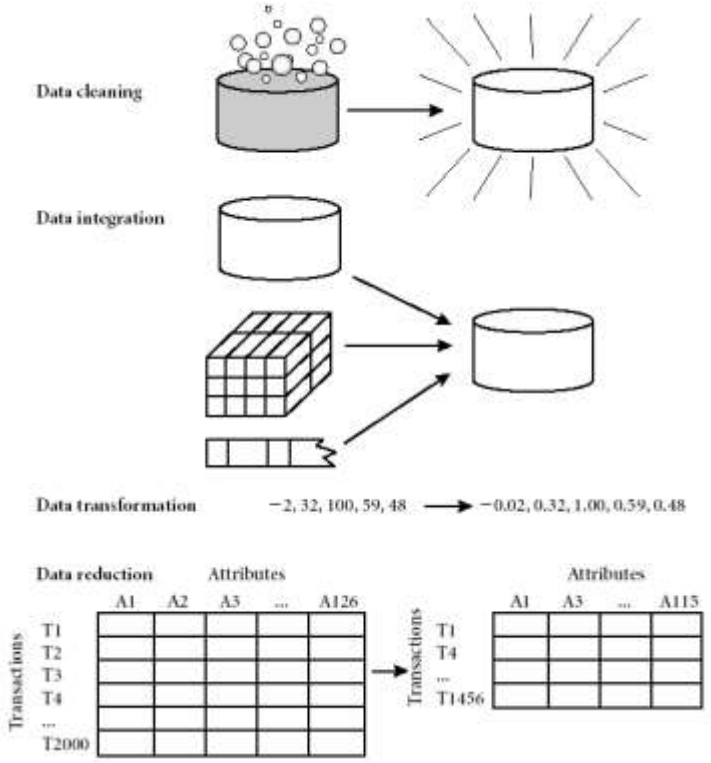


Figure 3.25 Selecting an interesting subset of records for further investigation.

Referinta figura: D. Larose, *Discovering Knowledge in Data. An Introduction to Data Mining*, John Wiley & Sons, 2005.

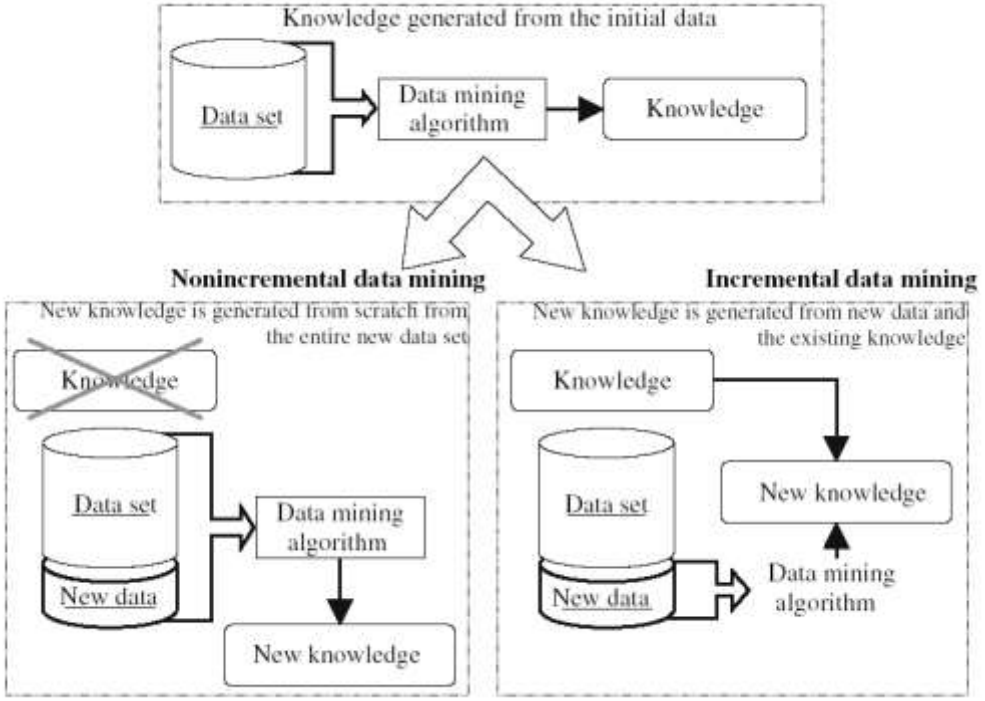
Preprocesarea datelor



Referinta figura: J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, 2nd Edition, Morgan Kaufmann, 2006.

Date Dinamice

- Datele sunt de obicei dinamice, adica noi obiecte si/sau atribute pot fi adaugate iar altele scoase sau inlocuite
- In acest caz, ar trebui ca si algoritmii de data mining sa evolueze in timp, ceea ce inseamna ca (,) cunostintele extrase trebuie modificate in timp
- Una dintre cele mai mari provocari pentru metodele de data mining este combinarea cunostintelor vechi cu cele extrase din datele noi



Referinta figura: K. Cios, W. Pedrycz, R. Swiniarski, L. Kurgan, *Data Mining. A Knowledge Discovery Approach*, Springer, 2007.