## UNIVERSITATEA POLITEHNICA DIN BUCURESTI FACULTATEA DE AUTOMATICA SI CALCULATOARE

Data Mining

|  |  |
| --- | --- |
|  | Enciu George Haralambie  355AA |

Cuprins

[Data Mining 3](#_Toc219662902)

[Analiza Componentelor Principale si Analiza Factoriala 5](#_Toc219662903)

[Analiza clasificarii (*Cluster Analysis*) 7](#_Toc219662904)

[Regresia Liniara Multipla si Regresia Logistica Binara 9](#_Toc219662905)

[Analiza de Discriminant 11](#_Toc219662906)

# Data Mining

*Data mining*, cunoscut si ca „descoperirea cunostintelor în baze de date mari” este un instrument modern si puternic al TI&C (Tehnologia Informatiei si Comunicatiilor), instrument ce poate fi folosit pentru extragerea unor informatii utile dar înca necunoscute. Acest instrument automatizeaza procesul de descoperire a unor relatii si combinatii în datele brute, iar rezultatele gasite ar putea fi încadrate într-un sistem automat de suport a deciziei.

*Data mining* a fost folosit pâna acum în afaceri de catre organizatii comerciale de succes în scopul de a obtine avantaje critice în competitia lor. Se banuieste ca în viitorul apropiat acest instrument va fi folosit pentru prelucrarea bazelor de date uriase, ca de exemplu dosarele computerizate ale pacientilor, la nivel national. De fapt, chiar si în prezent, prin identificarea procedurilor medicale ce au tendinta de a se grupa, prin *data mining* putem prezice care pacienti vor folosi noile strategii de îngrijire a sanatatii, putem defini modele de comportare ale pacientilor de risc, putem identifica fraudele.

Metodele *data mining* provin din calculul statistic clasic, din administrarea bazelor de date si din inteligenta artificiala. Ele nu înlocuiesc metodele traditionale ale statisticii, ci sunt considerate a fi extinderi ale tehnicilor grafice si statistice. Deoarece softului îi lipseste intuitia umana (pentru a face recunoasterea a ceea ce este relevant de ceea ce nu este), rezultatele metodelor *data mining* vor trebui supuse în mod sistematic unei supravegheri umane.

Structura tipica de date potrivita pentru *data mining* contine observatiile (cazurile, de exemplu referitoare la pacienti) plasate pe linii iar variabilele plasate pe coloane. Domeniile sau intervalele de valori pentru fiecare variabila vor trebui sa fie definite precis, evitându-se cât mai mult posibil exprimarile vagi. Formatul cu linii (ce corespund observatiilor sau cazurilor) si coloane (ce corespund variabilelor), similar cu cel al unui fisier *spreadsheet*, este obligatoriu pentru *data mining*. Pregatirea datelor pentru a le fi aplicate instrumentele din *data mining* consta în:

a) Preprocesare – eliminarea datelor care nu sunt necesare,

– verificarea consistentei (unitatilor de masura),

– detectarea datelor eronate si eliminarea lor,

– eliminarea valorilor extreme (*outliers*);  
b) Integrarea datelor – combinarea unor variabile;  
c) Transformarea variabilelor – prin standardizare,  
– prin trecere la scala logaritmica;  
d) Separarea bazei de date în trei categorii de date:

1) Categoria pentru antrenament,

2) Categoria pentru validare,

3) Categoria pentru testare.

e) Folosirea statisticilor descriptive clasice simple: media, mediana, amplitudinea,

abaterea standard, cuartilele;

f) Folosirea diagramelor simple: histograme ale frecventelor, *box plot*-uri, diagrame cu

bare, diagrame radiale (rozeta).

Sa ne reamintim ca:

– Histogramele frecventelor prezinta distributia valorilor variabilelor continue;

– *Box plot*-urile sumarizeaza (vizual) mai multe aspecte importante ale unei variabile

continue (mediana, cuartilele, extremele);

– Diagramele cu bare prezinta diferentele între diversele grupuri, pentru variabile

categoriale, calitative;

– Diagramele radiale compara între ele nivelurile claselor unei variabile calitative.

Metodele *data mining* „de învatare” pot fi grupate în doua categorii: cele nesupervizate, respectiv cele supervizate.

Metodele de învatare nesupervizate includ urmatoarele:

1. Analiza Componentelor Principale (ACP). Scopul sau este cel de a reduce dimensionalitatea datelor multi-variate prin „integrarea” variabilelor corelate, transformând liniar variabilele initiale în variabile necorelate între ele.
2. Analiza Factoriala (*Factor Analysis*). Scopul sau este cel de a extrage un numar mic de factori ascunsi care explica cea mai mare parte a variabilitatii comune si determina corelatiile observate între datele initiale.
3. Analiza Clasificarii (*Cluster Analysis*). Scopul sau este cel de a grupa cazurile (observatiile) în clustere (grupuri, categorii).

Principalele metode de învatare supervizata sunt urmatoarele:

1) Regresia Liniara Multipla. Scopul sau este cel de a descrie asociatiile între doua seturi de variabile, prin intermediul unor formule liniare. Aceste formule sunt folosite ulterior pentru a prezice valorile unor variabile continue „de raspuns” odata ce sunt cunoscute valorile tuturor variabilelor „predictor”.

2) Regresia Logistica. În acest tip de regresie „raspunsul” este o variabila binara sau ordinala (nu una continua).

3) Retelele Neurale. Sunt folosite pentru clasificare. Totusi, o retea neurala contine de obicei mult mai multi parametri decât un model statistic clasic, este nevoie de mult mai mult timp de antrenare, iar rezultatele obtinute nu pot fi interpretate cu usurinta.

4) Analiza Functiilor Discriminante (sau Analiza Discriminarii). Aceasta metoda este folosita pentru a determina care dintre variabilele predictor discrimineaza cel mai bine între mai multe grupuri care sunt formate natural.

Modelele identificate de o metoda de *data mining* vor putea fi transformate în cunostinte, însa dupa o validare corespunzatoare; apoi, cunostintele vor putea fi folosite ulterior pentru a fundamenta luarea deciziilor.

În continuare va fi prezentata esenta principalelor metode de *data mining*, începând cu cele nesupervizate.

# Analiza Componentelor Principale si Analiza Factoriala

Din cauza dificultatii evidente a vizualizarii spatiilor multi-dimensionale având dimensiuni *p* 4, ACP este folosita cel mai mult pentru a reduce pe cât posibil dimensionalitatea a *p* variabile la doar doua sau trei dimensiuni. Aceasta metoda sumarizeaza variabilitatea initiala a datelor privind cele *p* variabile initiale în câteva componente necorelate între ele, numite componentele principale. Fiecare componenta principala este extrasa ca o combinatie liniara de variabilele initiale. Metoda consta în extragerea celui mai mic numar de componente care preiau cea mai mare parte a variantei datelor initiale, adica în sumarizarea datelor initiale cu o pierdere minima de informatie”.

Prima componenta principala extrasa este acea combinatie liniara de variabile care preia maximul posibil din varianta datelor initiale. Cea de-a doua componenta principala preia mai putina varianta, si asa mai departe. Daca primele câteva componente principale preiau 80% sau mai mult din varianta datelor initiale, atunci scopul reducerii dimensionalitatii va fi atins.

Calculele sunt facute prin algoritmi ai algebrei liniare, aplicati fie matricei dreptunghiulare a datelor originale, fie matricei patratice a coeficientilor de corelatie. În teoria matematica ce fundamenteaza acesti algoritmi se foloseste termenul de **valoare proprie** (*eigenvalue*) a unei matrice. Valorile proprii masoara cantitatea de varianta „explicata” de fiecare componenta principala. Ele descresc odata cu indexul componentei, prima componenta principala având valoarea proprie maxima. Pe de alta parte, suma valorilor proprii este egala cu *p* (numarul variabilelor initiale). De obicei sunt retinute doar componentelor principale ce au valorile proprii mai mari decât 1. Valorile proprii exprima „importanta” componentelor principale.

O diagrama în care valorile proprii sunt prezentate ca linii verticale sau ca puncte deasupra numerelor naturale (care reprezinta componentelor) este numita „în coborâre” (*scree plot*). Un exemplu este prezentat în figura alaturata. Notiunea de **scor** (al unei observatii) poate fi înteleasa daca interpretam observatiile ca vectori într-un spatiu *p*-dimensional al variabilelor. În ACP acest spatiu este înlocuit cu un altul, cel al componentelor principale. Scorurile sunt exact vectorii ce reprezinta observatiile în acest nou spatiu.

În sfârsit, **încarcarile** (*loadings*) sunt coeficientii de corelatie între coloanele-scor si variabilele originale. Extrem de important este studiul coeficientilor de corelatie (*loadings*) dintre variabilele initiale si primele doua componente principale. Coeficienti puternici arata ca variabilele corespunzatoare pot fi considerate responsabile pentru variatia datelor. Din contra, daca o variabila nu se coreleaza cu nici o componenta principala, sau se coreleaza cu componentele ce au valori proprii mici, aceasta sugereaza ca variabila în cauza are o contributie minora la varianta setului de date. Astfel de variabile „neimportante” vor fi eliminate, în scopul de a simplifica analiza de ansamblu.

În unele situatii avem motive sa credem ca variabilele masurate sunt corelate între ele deoarece ele sunt influentate de unul sau mai multi factori necunoscuti, numiti factori ascunsi. Analiza Factoriala este o tehnica al carui scop este extragerea unui numar mic de factori ascunsi care sunt responsabili pentru corelatiile între variabile. Principalul rezultat al acestei tehnici, aplicata fie plecând de la matricea datelor, fie plecând de la matricea de corelatie, consta în gruparea variabilelor în asa fel încât variabilele influentate de un anumit factor sunt corelate mai puternic între ele decât cu variabilele influentate de alti factori.

Initial se accepta ca toate variabilele masurate ar putea fi asociate cu orice factor ascuns. Dupa aplicarea tehnicii, fiecare variabila masurata va fi exprimata ca o combinatie liniara (ponderata) de câtiva factori ascunsi.

În implementarea metodei AF se întâlneste termenul de comunalitate (*communality*).

**Comunalitatea** unei variabile masurate este exact proportia din varianta ei ce poate fi atribuita factorilor ascunsi. Evident, comunalitatea este un numar între 0 si 1, iar valorile apropiate de 1 indica faptul ca variabila noastra este „bine explicata” de factorii ascunsi.

**Încarcarile factorilor** (*factor loadings*) sunt coeficientii de corelatie între variabilele masurate si factorii ascunsi. Încarcari mai mari decât 0.7 în valoare absoluta sunt considerate semnificative. Încarcarile semnificative pentru un factor ascuns dat ar putea fi folosite pentru a obtine o interpretare a acelui factor.

Rezultat de AF obtinut cu *Statistica*

Pentru determinarea numarului de factori ascunsi se poate folosi un *scree plot*, sau poate fi ales un prag pentru comunalitati, la fel ca în metoda componentelor principale.

Câteodata, pentru a putea obtine factori care nu au multe încarcari „mici” (nesemnificative) – prin urmare pentru a simplifica interpretarea factorilor – este efectuata o „rotatie”; aceasta este cunoscuta sub numele de metoda Varimax.

Sa prezentam, în figura de mai sus, un rezultat tipic al tehnicii Analizei factoriale. Este o diagrama planara în care variabilele masurate CA, CE, IM, RI, SN, TV si UR sunt înlocuite prin puncte din interiorul cercului unitate. Cei doi factori ascunsi sunt înlocuiti prin punctele identificate ca F1 si F2. Rezultatele numerice, dupa efectuarea unui Varimax, sunt prezentati în tabelul de sub diagrama. Variabilele CE, IM, TV si UR pot fi grupate într-un grup si reprezentate de factorul F1; pe diagrama este clar ca variabilele CE, IM, TV, UR sunt puternic (negativ) corelate cu F1 (si între ele). Astfel Factorul 1 reprezinta „ceva ce variabilele CE, IM, TV, UR au în comun”. Variabilele CA si RI pot fi grupate în alt grup si înlocuite prin F2. Este clar ca CA este puternic corelat pozitiv cu F2 iar RI este puternic corelat negativ cu F2. (Am putea spune ca variabilele CA si RI au tendinta de a varia în directii opuse.) Pe de alta parte, variabilele MI si SN sunt „neutre” în raport cu ambii factori. Factorului 1 i se atribuie mai mult de 54% din varianta totala, iar ambilor factori li se atribuie mai mult de 54+18 = 72% din varianta totala.

Analiza Factoriala este similara cu Analiza Componentelor Principale, ambele având ca scop reducerea numarului (mare de) variabile masurate într-un numar mai mic de factori ascunsi respectiv componente principale. Ele difera totusi ca utilitate. În AF numarul (mic) de factori are identificat în asa fel încât sa explice de ce variabilele masurate sunt corelate între ele. Din contra, în ACP componentele principale sunt identificate în asa fel încât sa preia cât mai mult din varianta prezenta în datele masurate.

# Analiza clasificarii (*Cluster Analysis*)

Aceasta metoda statistica este folosita pentru a grupa date multi-dimensionale (adica „puncte” ce reprezinta cazuri sau observatii) în grupe (*clusters*) definite algoritmic. Aceasta metoda este utila pentru sumarizarea unor cantitati mari de informatie, fiecare grupa reprezentând mai multe puncte având caracteristici similare. Clusterele distincte nu se suprapun (adica sunt disjuncte).

De fapt, analiza clasificarii consta dintr-o colectie de algoritmi ce exploateaza mai multe euristici fundamentate în principal pe experienta noastra „vizuala” în gruparea punctelor în „nori de puncte”.

În general, pentru a putea folosi un algoritm de clasificare, este nevoie de precizarea:

a) Unei distante între punctele unui spatiu multidimensional. Cele mai cunoscute distante

între punctele *A* (*a*1,*a*2,...) si *B* (*b*1,*b*2,...) sunt

* Euclidiana
* Manhattan
* Pearson

b) O strategie de alegere a punctului reprezentativ (adica a „centrului”) pentru orice grupare de puncte. Cei mai multi oameni au tendinta de a alege media aritmetica (adica „centrul de greutate”).

**c)** O distanta între doua grupe de puncte. Cele mai folosite asemenea distante iau în considerare distanta între puncte aleasa anterior; acestea sunt (a se vedea figura de mai sus):

– Distanta între „centre” (în algoritmul Ward);

– Distanta între cei mai apropiati vecini (optiune cunoscuta ca „*single linkage*”);

– Distanta între cei mai departati vecini (optiune cunoscuta sub numele de „*complete*

*linkage*”).

Odata ce au fost efectuate alegerile, un **algoritm ierarhic de clasificare** va functiona în felul urmator:

Pasul 1. Fiecare punct este considerat ca grupa separata (de 1 punct).

Pasul 2. Cele mai apropiate doua grupe sunt amalgamate într-o grupare mai mare.

Acest pas este repetat pâna când toate punctele au fost grupate într-o grupare finala (care „contine totul”). Schema de amalgamare este reprezentata diagramatic printr-o dendrograma (*hierarchical tree plot*) – a se vedea figura de mai jos pentru un exemplu).

Pasul 3. Se aplica o procedura „de taiere” asupra dendrogramei; în acest fel se identifica numarul „obiectiv” de grupe (*clusters*), apoi componenta fiecareia. (Aflati detalii despre algoritmul lui Ward si despre metoda poligonului lui Newton, ca procedura „de taiere”.)

Sa facem observatia ca în clasificarea ierarhica nu se permite nici un fel de alta suprapunere a doua grupe ce exceptia incluziunii unei grupe în cealalta. În plus, din cauza numarului mare de distante care ar trebui calculate, acesti algoritmi nu sunt adecvati pentru seturi mari de puncte.

Prin contrast, în **clasificarea disjuncta** numarul de grupe va trebui ales dinainte. Apoi:

Pasul 1. Toate punctele sunt asignate arbitrar unor grupe.

Pasul 2. Unul dintre puncte este reasignat altei grupe, luând în considerare similaritatea sa cu punctele acelei grupe. Acest pas este repetat pâna când toate punctele sunt asignate „optim”.

Algoritmii de clasificare sunt mai eficienti pentru seturi mari de puncte. (Dar sa ne aducem aminte ca au si un neajuns: numarul de grupe este ales arbitrar!) Cel mai cunoscut algoritm de clasificare disjuncta este cel al k-mediilor (*k-means*). În acesta reprezentantul fiecarei grupe este obtinut din media punctelor (cazuri, observatii) care sunt asignate acelei grupe. Initial este selectat un set de puncte, numite „semintele grupelor” (*cluster seeds*) ca o prima alegere a centrelor grupelor. Apoi, la fiecare pas, se executa maximizarea distantei globale între grupele distincte, în asa fel încât sa se reduca varianta din cadrul grupelor, iar centrele sa se deplaseze spre o pozitie stabila. Algoritmul se opreste atunci când nu se mai detecteaza nici o modificare în pozitia centrelor. Sa facem observatia ca variantele variabilelor au o influenta puternica asupra distantelor, în special asupra celor Euclidiene. De exemplu, atunci când se calculeaza distante, datele în domeniul 100-1000 sunt practic decisive comparativ cu cele din domeniul 0.01-0.1.

# Regresia Liniara Multipla si Regresia Logistica Binara

Aceasta metoda de învatare supervizata este cea mai larg folosita astazi. Este utilizata pentru a studia asocierea dintre doua seturi de variabile: primul set {*X*1, *X*2,...} contine

variabilele predictor (numite si „independente” sau „explicative”), celalalt set {*Y*1,*Y*2,...}

contine variabilele raspuns (numite si „dependente”). Asocierea dintre cele doua seturi este descrisa prin formulele liniare

în care *bj*0,*bjk* sunt parametrii de regresie.

Daca aceasta formula este „suficient de buna”, ea va putea fi folosita pentru a prezice valorile *yj* ale variabilelor *Yj* odata ce sunt cunoscute valorile *xk* ale variabilelor independente *Xk* .

Parametrii de regresie – ce apar în formula de mai sus – sunt estimati luând în considerare datele de antrenament cunoscute prin folosirea criteriului celor mai mici patrate:

adica prin minimizarea sumei patratelor diferentelor dintre valorile observate *y* ale variabilelor-raspuns si valorile corespunzatoare ale variabilelor predictor.

Sa consideram cazul particular al unei singure variabile-raspuns *Y*.

Primul obiectiv al metodei regresiei liniare multiple îl constituie obtinerea unui hiperplan care se potriveste optimal „norului” de puncte format de datele de antrenament; acest scop este atins prin calcularea estimatiilor *β*0,*βk* pentru parametrii *b*0,*bk* .

Formula de calcul finala este

în care *β*0 este numit **interceptul**, iar *βk* sunt numiti **coeficientii de regresie**.

(De obicei estimarile *βk* sunt notate *bk* , „caciula” având exact întelesul de „estimare”.)

Numerele *βk* exprima cu cât se modifica raspunsul *Y* atunci când predictorul *Xk* suporta o modificare de o unitate, iar ceilalti predictori nu-si modifica valoarea. (Atentie, are sens sa comparam între ei doi coeficienti de regresie *βk* si *βl* doar daca predictorii corespunzatori *Xk* si *Xl* sunt masurati cu aceeasi unitate de masura!)

O notiune importanta în folosirea metodei RLM este cea de reziduu. Reziduurile sunt exact diferentele dintre valorile *y* observate si valorile-raspuns *y* ce corespund valorilor predictor *xk* observate. Aceste diferente sunt calculate folosindu-se estimarile *β*0,*βk* , mai precis astfel

unde *i* „numara” observatiile din setul de date de antrenament.

În aplicarea concreta a metodei regresiei liniare multiple se fac urmatoarele presupuneri (considerate implicit ca adevarate):

1) Reziduurile *r*(*i*) obtinute din regresie sunt realizari ale unor variabile aleatoare normale de medie 0 si aceeasi varianta, variabile care sunt si independente între ele;

2) Nu este prezenta multi-colinearitatea, adica nici o variabila predictor nu este o

combinatie liniara de celelalte variabile predictor.

În figura urmatoare sunt prezentate doua exemple de reziduuri care nu valideaza metoda RLM. În asemenea situatii, o abordare posibila ar consta în corectarea neliniara a variabilelor.

Exemple de reziduuri care nu valideaza RLM

 

Defectul major al RLM consta în faptul ca nu putem fi siguri asupra mecanismului relatiei cauzale dintre predictori si raspunsuri. Chiar si relatiile de regresie foarte semnificative nu implica (logic) deloc existenta unor relatii cauza-efect! Cu toate acestea, metoda RLM este des folosita, iar formulelor semnificative de regresie sunt creditate ca adevarate în activitatile de predictie.

Rezultatul unei RLM obtinut prin folosirea setului de date de antrenament, cu alte cuvinte formula de calcul, ar putea fi validat în urma aplicarii formulei asupra altui set de date si constatarii potrivirii ei. Pentru validare, graficele obtinute din setul de antrenament si din setul de validare ar trebui sa prezinte trasaturi similare.

Sa ne reamintim ca metoda regresiei liniare presupune ca variabilele – atât cele predictor cât si cele raspuns – sunt de tip continuu. Prin contrast, Regresia Logistica permite ratarea altor tipuri de variabile. Metoda Regresiei Logistice Binare se aplica atunci când variabila de raspuns este de tip Boolean (adica are doar doua valori: adevarat/fals, sau da/nu, sau 0/1).

Functia pas, definita de formula urmatoare, transforma numerele reale în valori Booleene:

# Analiza de Discriminant

Întro situatie tipica, un medic curant ar putea lua în considerare o serie de date privind trecutul pacientului, date pe care el le considera importante pentru ca ele indica daca pacientul are sanse de a se vindeca complet, sau doar partial, sau deloc. Scopul sau initial este de fapt de a construi un model predictiv de discriminare între grupuri, bazat pe variabile predictor.

Un asemenea model ar putea fi folosit pentru:

– Investigarea diferentelor dintre categoriile de observatii,

– Discriminarea efectiva între categorii,

– Identificarea acelor variabile predictor care sunt discriminatoare.

Într-o alta situatie tipica, un specialist ar putea cauta de exemplu criterii de discriminare între tipurile de celule canceroase. În ambele situatii de mai sus datele initiale sunt grupate natural în câteva categorii (trei în primul caz). Metodele Analizei de discriminant presupun ca toate variabilele predictor sunt continue; din contra, se presupune ca variabila de raspuns indica gruparea, categoria careia îi va apartine un nou caz. O prima metoda este cea numita analiza de discriminant progresiva în pasi (*forward stepwise discriminant analysis*). Pe scurt, aceasta consta în urmatoarele: initial nici o variabila predictor nu este inclusa în model. Apoi, la fiecare pas, sunt examinate toate variabilele predictive ramase, iar cea care contribuie cel mai mult la discriminarea între grupuri va fi selectata si inclusa în model.

În metoda duala, asa-numita analiza de discriminant regresiva în pasi (*backward stepwise discriminant analysis*), procedura este urmatoarea: initial toate variabilele predictor sunt incluse în model iar apoi, la fiecare pas, aceea care contribuie cel mai putin la predictia apartenentei corecte la grupare este eliminata. În model vor ramâne în final doar variabilele „importante”, acelea care contribuie cel mai mult la discriminarea între grupe.

În ambele metode de selectie a variabilelor, prezentate mai sus, procedura pas cu pas este ghidata de o valoare care este calculata pentru fiecare variabila predictor. Aceasta, cunoscuta ca valoarea F, indica semnificatia statistica a acelei variabile pentru discriminarea între grupe.

La fiecare pas, se ia în considerare si semnificatia predictorului anterior introdus în model. Variabila care contribuie cel mai putin la „puterea discriminatorie” a modelului este eliminata, iar apoi variabila care nu este în model si contribuie cel mai mult la discriminare va fi introdusa în model. Atunci când toate variabilele din model nu vor mai suferi modificari, procedura pas cu pas este oprita. Pentru a forta oprirea procedurii, pot fi fixate dinainte praguri pentru valoarea F (identificate ca *F-to-enter* si *F-to-remove* în figura urmatoare).

În cazul cel mai simplu (doar doua categorii-grupe), Analiza de Discriminant este practic echivalenta cu Analiza de regresie multipla, modelul constând în acest caz dintr-o singura formula liniara

în care *βk* sunt coeficientii de regresie.

Trebuie sa fim constienti de faptul ca indicele celui mai mare coeficient de regresie (în valoare absoluta) identifica acea variabila care contribuie cel mai mult la predictia apartenentei la categoria-grupa. Formula de mai sus descrie **functia de discriminare**.

În cazul existentei mai multor categorii, vor fi obtinute mai mult de o functie de discriminare. De exemplu, în cazul existentei a trei categorii, o prima functie de discriminare ar putea discrimina între categoria 1 si categoriile 2-3 combinate, iar a doua functie de discriminare va discrimina între categoria a 2-a si 3-a.

Modelele de discriminare între categorii vor trebui validate pe baza unor date noi, altele decât cele pe baza carora au fost construite.

Metodele de selectie în pasi nu dau întotdeauna cele mai bune rezultate. Unele motive ar fi: a) în procesul de selectie nu sunt luate în considerare relatiile între variabilele care înca n-au fost selectate; b) nu se iau în considerare diferentele între marimea grupelor. (Aceasta marime a grupelor influenteaza probabilitatile *a priori*, cu alte cuvinte probabilitatea ca un caz nou sa apartina unei grupe particulare.)

Datorita cresterii influentei Tehnologiei Informatiei si Comunicatiilor în lumea moderna, recent au fost imaginate metode noi în *Data Mining*. Printre aceste metode, aparute în lumea finantelor, se afla:

– Aplicatiile depozitului de date (*data warehousing*), si

– Analiza asocierilor din cosul de piata (*market basket association analysis*).

În aplicatiile depozitului de date obiectivele sunt: utilizarea completa a surselor de date, extragerea datelor din diferite locatii, integrarea datelor în baze de date multidimensionale) de înalta calitate, optimizarea stocarii datelor. În acest context sa prezentam doar o idee specifica, anume cea de metadata. O **metadata** contine, pe lânga datele ca atare, si informatii despre acestea: sursa lor, transformarile la care au fost supuse, intervalul de timp etc.

Obiectivul principal al analizei asocierilor din cosul de piata consta în identificarea produselor si serviciilor pe care clientii (de exemplu ai unui supermarket) le achizitioneaza împreuna. Cunoasterea asocierilor între produsele si serviciile asociate ajuta managerii în luarea unor decizii bune în tacticile de marketing si vânzare.

Este evident ca aceste metode noi ar putea fi aplicate pentru a îmbunatati calitatea sistemului de sanatate, cel putin la nivelul spitalului. (Imaginati-va beneficiile unei analize a asocierilor din cosul de piata, efectuate pe datele medicale la nivel national.) datele ca atare, si informatii despre acestea: sursa lor, transformarile la care au fost supuse, intervalul de timp etc.

Obiectivul principal al analizei asocierilor din cosul de piata consta în identificarea produselor si serviciilor pe care clientii (de exemplu ai unui supermarket) le achizitioneaza împreuna. Cunoasterea asocierilor între produsele si serviciile asociate ajuta managerii în luarea unor decizii bune în tacticile de marketing si vânzare. Este evident ca aceste metode noi ar putea fi aplicate pentru a îmbunatati calitatea sistemului de sanatate, cel putin la nivelul spitalului. (Imaginati-va beneficiile unei analize a asocierilor din cosul de piata, efectuate pe datele medicale la nivel national.)