Data mining - de la sisteme de baze de date la sisteme cu baze de cunostinte

Prof.dr. Constanta BODEA, lect. Cristina IONITA,
asist. Anton CRETU, prep. Petrisor OPREA
Catedra de Informatica Economica, A.S.E. Bucuresti

Una dintre cele mai importante surse de achizitionare a cunostintelor o constituie, în prezent, bazele de date ale întreprinderii. Utilizarea instrumentelor de achizitionare si utilizare a cunostintelor în sistemele de baze de date a devenit o necesitate pentru toate organizatiile care folosesc un volum mare de date, colectate din diferite surse. Aceste instrumente au la baza o serie de metode, de la metodele statistice de interogare si raportare clasice pâna la metodele de învatare automata. Articolul prezinta câteva din aceste metode, punând în evidenta caracteristicile lor principale precum si avantajele si dezavantajele în utilizare. Pentru doua dintre aceste metode, analiza multidimensionala si calculul neuronal sunt prezentate rezultatele obtinute în cadrul unui studiu caz pentru domeniul vânzarilor.

Cuvinte cheie: *achizitionarea cunostintelor, utilizarea cunostintelor, data mining, sisteme de baze de date multidimensionale, calcul neuronal.*

1. Metode de achizitionare si utilizare a cunostintelor în sistemele de baze de date

Infrastructura informationala a organizatiilor este în continua crestere si tinde sa atinga un grad de maturitate caracterizat de un anumit nivel al calitatii datelor si al facilitatilor de acces la date. Pe masura ce infrastructura informationala se maturizeaza, creste necesarul de informatie relevanta, ceea ce reclama gasirea unor solutii de crestere a valorii datelor [3]. În forma în care sunt culese, datele au un nivel scazut de relevanta. Pentru a le utiliza, datele trebuie pregatite si prelucrate prin selectii, proiectii, reducerea dimensiunii, extragerea de pattern-uri si modele. Prin aceste operatii obtinem *cunostinte*, care sunt transmise managerilor pentru a fi folosite în procesul de luare a deciziilor sau sunt utilizate chiar de sistem pentru a ajunge la anumite concluzii relevante pentru manageri. Sistemele de baze de date (BD) tind sa devina sisteme de achizitionare si utilizare a cunostintelor (AUC).

Cei mai importanti pasi pentru realizarea AUC în domeniul BD sunt:

a) regasirea si raportarea datelor relevante din BD;

b) prelucrarea datelor (selectie, proiectie, reducerea dimensiunii);

c) transformarea datelor (totalizare, normalizare);

d) extragerea de pattern-uri si modele;

e) utilizarea pattern-urilor si modelelor predeterminate pentru a ajunge la cunostinte relevante (selectarea modelului, testarea si validarea modelului, dezvoltarea modelului).

Pasii a)-d) sunt referiti, în general, drept etape ale procesului de achizitionare a cunostintelor, în timp ce pasul e) reprezinta procesul de utilizare a cunostintelor. Pasul d) este denumit si procesul de extragere a cunostintelor din date (*data mining*).

Cele mai raspândite metode pentru realizarea AUC în sistemele de BD sunt urmatoarele: metodele statistice si de raportare (metodele OLAP); metodele neuronale; metodele de învatare automata; metodele bazate pe multimi aproximative (rough sets).

În figura 1 se prezinta trendul utilizarii acestor metode în sistemele de BD.

2. Stabilirea unei metrici pentru analiza comparativa a metodelor AUC

Evaluarea avantajelor utilizarii diferitelor metode de AUC poate fi realizata prin raportarea la cerintele domeniului de aplicabilitate si la resursele solicitate.

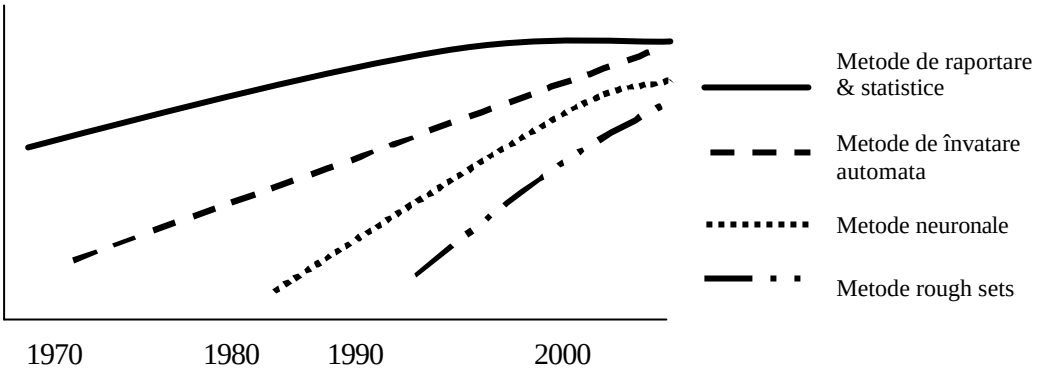

Fig. 1. Trendul utilizarii metodelor AUC

În tabelul 1 sunt prezentate dimensiunile analizei comparative între principalele metode de AUC, dimensiuni grupate în patru clase: (1) calitatea rezultatelor si necesarul de resurse, (2) cerinte de inginerie software, (3) restrictii logistice, (4) aria de aplicabilitate si restrictii de integrare [3].

Tabelul 1 - Analiza comparativa a principalelor metode de AUC

Dimensiuni	**Rezultate statistice si de raportare**	**Metode neuronale**	**Metode de învatare automata**	**Metode bazate pe multimi aproximative**
1) calitatea rezultatelor si necesarul de resurse				
acuratete	medie	mare	medie	medie
explicabilitate	medie	mica	medie	medie
viteza / fiabilitate	mare	mare	mare	medie
toleranta la zgomote	mica	mare	medie	mica
toleranta la raritatea datelor	mica	mica	mica	medie
toleranta la complexitate	medie	mare	medie	mare
2) cerinte de inginerie software				
flexibilitate	mare	mare	mare	mare
scalabilitate	medie	medie	mare	medie
compactare	mica	mare	medie	mare
usurinta în utilizare	mare	medie	medie	medie
3) restrictii logistice				
independenta fata de experti	medie	mare	medie	medie
usurinta în calcul	medie	medie	medie	mare
timp de dezvolare	mare	mediu	mediu	mediu
4) aria de aplicabilitate si restrictii de integrare				
pasii procesului de AUC	a+b+c	c+d+e	c+d	c+d
variabilitatea domeniilor de utilizare	mare	mare	medie	medie
migrarea datelor	mare	medie	mica	medie

Relevanta rezultatelor analizei comparative depinde de domeniul de aplicabilitate a metodelor, de instrumentele hardware si software utilizate pentru implementarea metodelor.

Pentru a fundamenta analiza metodelor AUC în sistemele de BD am realizat un studiu de caz pentru un sistem tranzactional al vânzarilor. În acest sistem, datele despre vânzari sunt organizate într-o baza de date relationala (BDR) a carei structura este redata în figura 2. Managerii au nevoie de o solutie informatica pentru descoperirea pattern-urilor de evolutie a vânzarilor si previzionarea vânzarilor pe diferite arii geografice, categorii de produse sau tipuri de clienti.

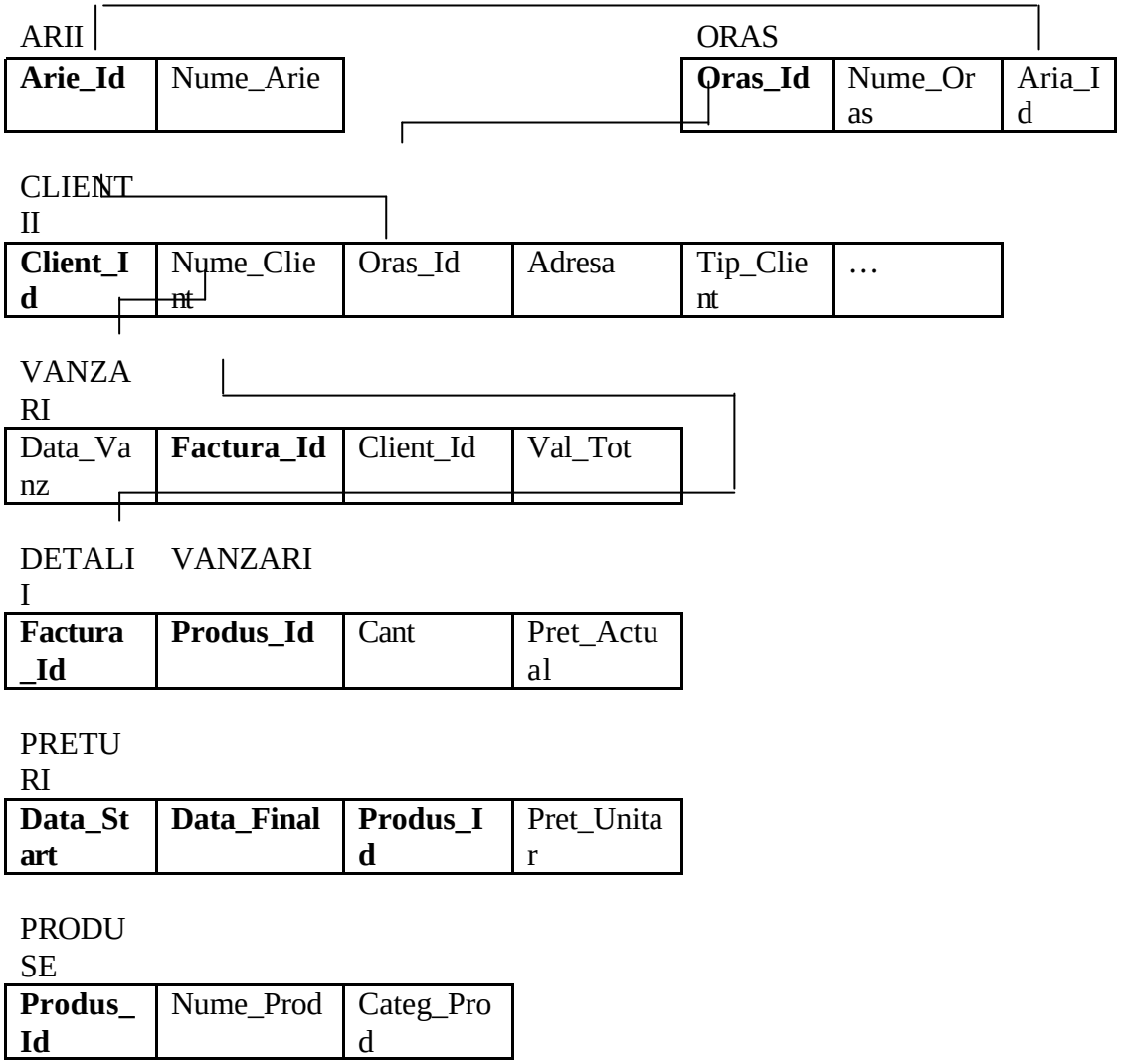

Fig. 2. Structura BDR pentru sistemul de vânzari

Pentru ilustrarea procesului de AUC din BDR pentru vânzari am ales doua metode: analiza multidimensionala si calculul neuronal. Urmatoarele sectiuni ale articolului prezinta rezultatele obtinute în analiza vânzarilor, utilizând ORACLE Express pentru analiza multidimensionala si toolbox-ul Neural Networks din MATLAB pentru calcul neuronal.

3. Analiza multidimensionala în sistemele de BD – Studiu de caz

Managerul de vânzari analizeaza activitatile economice în functie de factorii ("dimensiunile") care influenteaza evolutia activitatilor, cum ar fi: timpul, aria geografica, categoria de produs si tipul de client. Managerii nu gândesc în termenii relatiilor bidimensionale, plate, din BDR care presupun realizarea a numeroase jonctiuni pentru a obtine o imagine detaliata a vânzarilor care

sa surprinda toti factorii de influenta. Ei vizualizeaza datele despre vânzari sub forma de hipercub în care dimensiunile de analiza (timp, aria de vânzare, produsul, clientul) sunt muchiile hipercubului, iar datele despre vânzari (valoarea vânzarilor, cantitatea, pretul mediu) sunt celulele din interiorul hipercubului [9]. O solutie eficienta de stocare si prelucrare a hipercuburilor de date este oferita de bazele de date multidimensionale (BDM).

Pentru studiul de caz privind analiza vânzarilor s-a proiectat o BDM a carei structura este prezentata în figura 3. BDM are o tabela centrala de fapte care stocheaza informatiile privind valoarea vânzarilor, cantitatea si pretul mediu si mai multe tabele dimensiune pentru a stoca informatiile privind factorii care influenteaza evolutia vânzarilor. Unele din tabelele de dimensiune au o structura ierarhica (TIMP, PRODUS) iar altele sunt ierarhic legate între ele (CLIENT, ORAS, ARIA) pentru a putea realiza agregarea si descompunerea datelor în raport cu aceste dimensiuni.

Pentru a ilustra tehnicile de analiza OLAP pe o BDM am ales instrumentele din suita ORACLE Express, care ofera facilitati de analize statistice si raportare avansate (analize pe serii de timp, regresii, simulari, prognoze).

În procesul de descoperire a cunostintelor deosebit de eficiente sunt facilitatile oferite utilizatorilor de a sectiona hipercuburile de date si respectiv de a agrega datele din hipercuburi pe o singura dimensiune sau pe mai multe, operatii cunoscute sub numele de *drill down* si respectiv de *roll-up*. Astfel, managerul poate începe analiza cu vizualizarea agregata a datelor. Daca observa un fenomen si doreste sa descopere factorii care au influentat producerea acestuia, poate începe un proces de sectionare succesiva a datelor pe o singura dimensiune sau corelat pe mai multe dimensiuni.

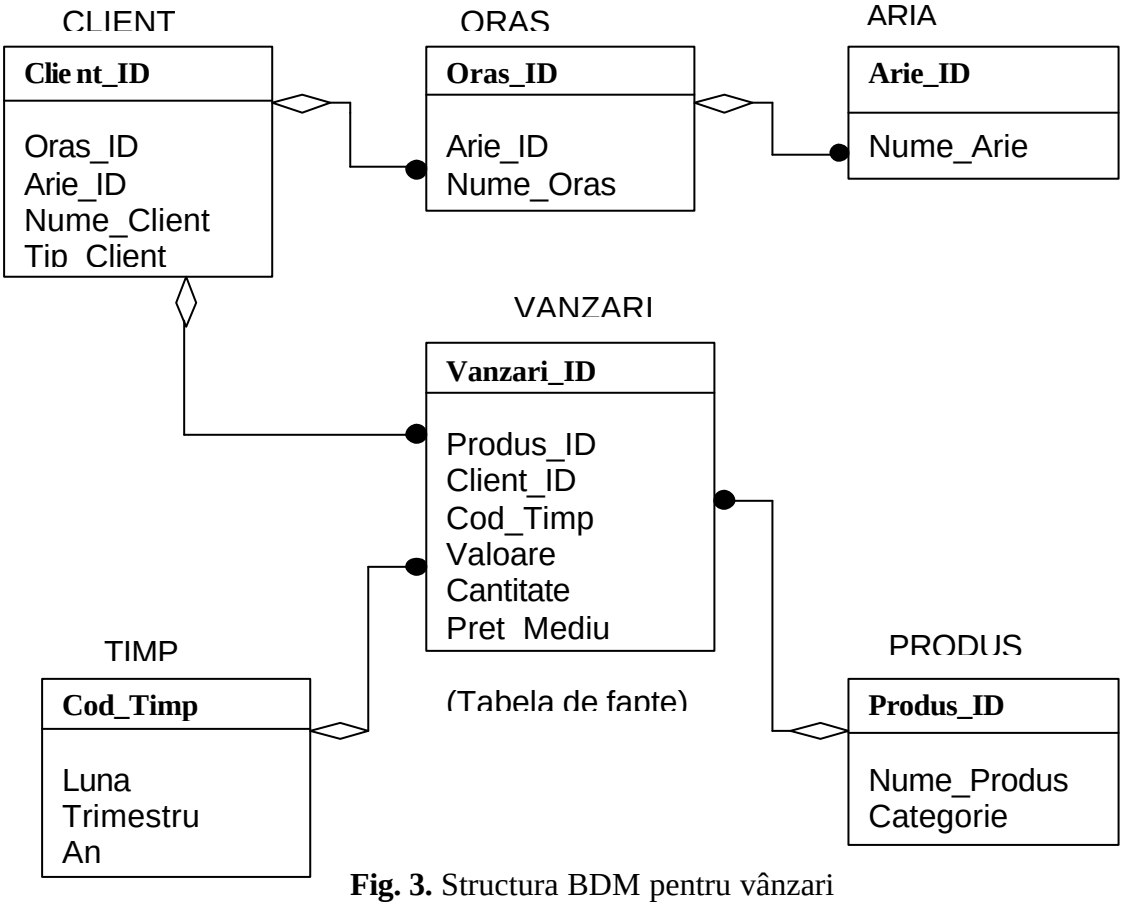

Fig. 3. Structura BDM pentru vânzari

De exemplu, analizând valoarea agregata a vânzarilor pe fiecare trimestru din ultimii cinci ani (1994 -1998) din figura 4 managerul poate sesiza vizual ca, desi tendinta generala este de crestere a vânzarilor, în trimestrul 1 si trimestrul 4 vânzarile sunt mai scazute. Sectionând datele privind valoarea trimestriala a vânzarilor pe categorii de produse (figura 5) managerul va sesiza ca produsele din categoria 3 prezinta o sezonalitate a vânzarilor, înregistrând o scadere a vânzarilor mai mica în trimestrul 4 si mai accentuata în trimestrul 1.

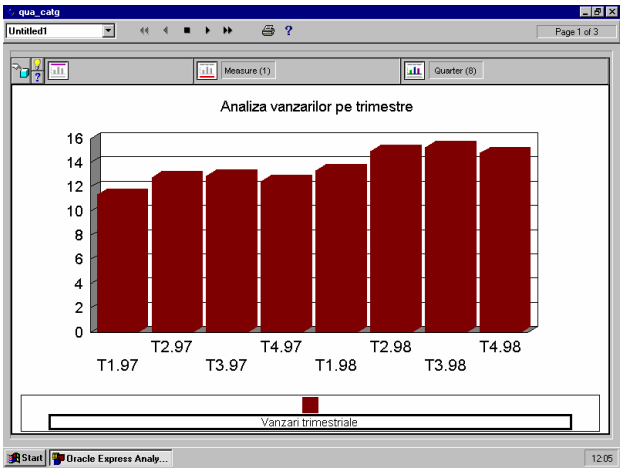

Fig. 4. Analiza vânzarilor pe trimestre

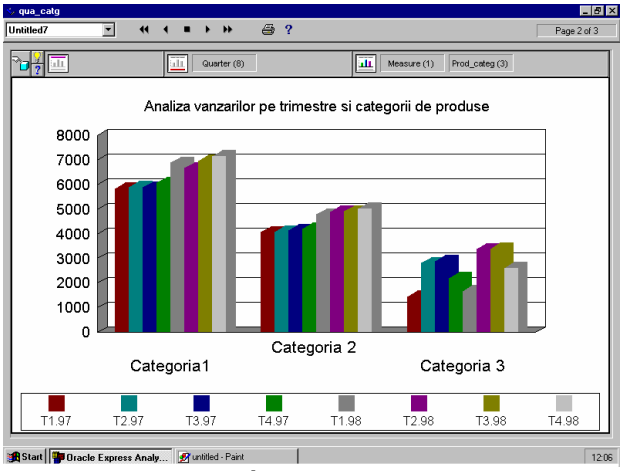

Fig. 5. Analiza vânzarilor pe trimestre si pe categorii de produse

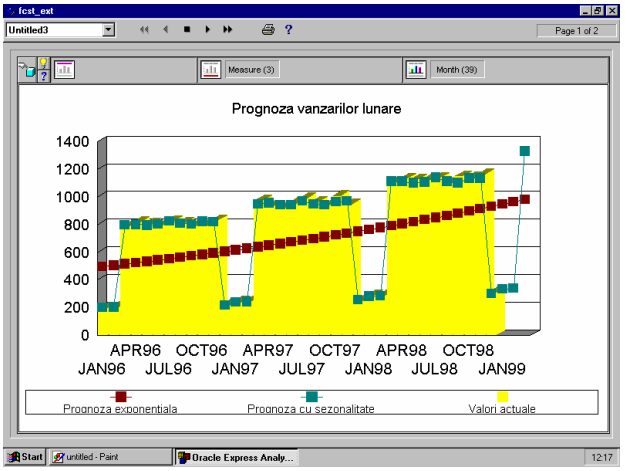

Fig. 6. Rezultatele procesului de prognoza pentru produsele din categoria 3

Prognoza datelor în Express presupune mai întâi identificarea vizuala a trendului de evolutie si apoi alegerea metodei de prognoza: liniara, exponentiala sau cu sezonalitate. Pentru prognoza vânzarilor pe primele trei luni din anul 1999, pe baza datelor privind vânzarile lunare din perioada 1996-1998 am folosit doua metode de prognoza: exponentiala si cu sezonalitate. Rezultatele celor doua metode de prognoza sunt prezentate comparativ în graficul din figura 6 numai pentru produsele din cate-goria 3. Se observa ca prognoza cu sezo-nalitate de tip Holt-Winters aproximeaza datele cu o exactitate mai mare. Rezultatele valorice ale prognozei sunt prezentate în tabelul 3.

4. Calcul neuronal în sisteme de baze de date – studiu de caz

Scopul propus este de a determina functiile care aproximeaza trendul vânzarilor pe cele 3 categorii analizate de produse, în vederea prognozarii vânzarilor lunare.

Pentru realizarea acestui scop am definit câte o arhitectura de retea neuronala feed-forward multinivel pentru fiecare din cele 3 categorii de produse, folosind algoritmul faster-backpropagation pentru antrenarea acestor retele. Fiecare retea este constituita din:

- un nivel de intrare, cu 5 unitati; 3 unitati de intrare semnifica vânzarile pentru categoria analizata de produse în ultimele 3 luni, iar 2 unitati de intrare semnifica vânzarile pentru categoria analizata de produse în aceeasi luna a ultimilor 2 ani;
- un nivel de iesire, cu un element, semnificând nivelul vânzarilor pentru categoria analizata de produse în luna curenta;
- un nivel intermediar (ascuns), continând 5 elemente de procesare. Numarul neuronilor de pe nivelul ascuns a rezultat în urma mai multor sesiuni de antrenare, luând în considerare timpul necesar antrenarii, precum si o aproximare cât mai exacta a trendului.

Pentru a realiza simularea evolutiei vânzarilor pentru categoriile analizate de produse am utilizat toolbox-ul Neural Networks din MATLAB. Datele necesare au fost exportate din baza de date în fisiere text, de unde au fost importate în MATLAB.

În faza de antrenare am utilizat pentru fiecare retea 36 seturi de antrenare, care contin ca outputuri dorite vânzarile pe 36 luni (din ianuarie 1996 pâna în decembrie 1998) plus cele 5 inputuri cu semnificatia mentionata anterior. Functia de activare a neuronilor este de tip logistica sigmoida, atât pentru neuronii de pe nivelul ascuns, cât si pentru neuronii de pe nivelul de iesire. Pentru a realiza o aproximare cât mai buna si mai rapida a trendului, datele din seturile de antrenare (atât inputurile cât si outputurile) au fost normalizate astfel:

$Xn_{ij} = (X_{ij} - Xmin_j) / (Xmax_j - Xmin_j)$, i=1,Q; j=1,R+S

unde: Q este numarul de seturi de antrenare; R este numarul de inputuri; S este numarul de outputuri; X_{ij} reprezinta datele preluate din seturile de antrenare; $Xmax_j$, $Xmin_j$ reprezinta marginile superioare si inferioare (determinate pe fiecare input si output); Xn_{ij} reprezinta datele normalizate.

Pe parcursul antrenarii se urmareste minimizarea erorii globale, calculata ca suma a erorilor (diferentelor) între outputurile dorite si outputurile obtinute din propagarea inputurilor în retea. Cele 3 retele au fost antrenate pâna când aceasta eroare a scazut sub o limita specificata. Numarul epocilor de instruire si eroarea globala pentru fiecare din cele 3 retele sunt prezentate în tabelul 2.

Tabelul 2.

Retea/Categorie de produs	Numar epoci	Eroare globala
Categoria 1	160000	0.0067
Categoria 2	160000	0.0064
Categoria 3	180000	0.0057

Evolutia erorii pentru cea de a 3-a categorie de produse este prezentata în figura 7, iar functia care aproximeaza trendul vânzarilor pentru categoria a 3-a de produse (rezultata în urma antrenarii retelei) este prezentata în figura 8.

În faza de simulare, pentru fiecare categorie de produse am utilizat ca seturi de test inputurile din seturile de antrenare, pentru a verifica daca aproximarea este corecta. Dupa aceasta am realizat câteva prognoze ale vânzarilor lunare, începând cu ianuarie 1999. Câteva din rezultatele simularilor sunt prezentate în tabelul 3.

5. Concluziile analizei comparative

Modelele multidimensionale de organizare a datelor ofera un suport eficient pentru realizarea prelucrarilor analitice de tip OLAP si pentru fundamentarea deciziilor. Sistemele OLAP ce implementeaza BDM sunt orientate catre decidenti si permit acestora

sa realizeze cu usurinta sectionari si grupari ale datelor supuse analizei si sa vizualizeze datele din diferite perspective pentru a descoperi factorii (dimensiunile) care influenteaza un anumit comportament. Sistemele OLAP ofera facilitati de analiza statistica avansate cum ar fi analize pe serii de timp, regresii, simulari, prognoze care au la baza modele predefinite. Instrumentele de tip OLAP precum ORACLE Express sunt rapide, usor de utilizat si gestioneza volume mari de date de ordinul megabytes si gigabytes într-un timp relativ scurt.

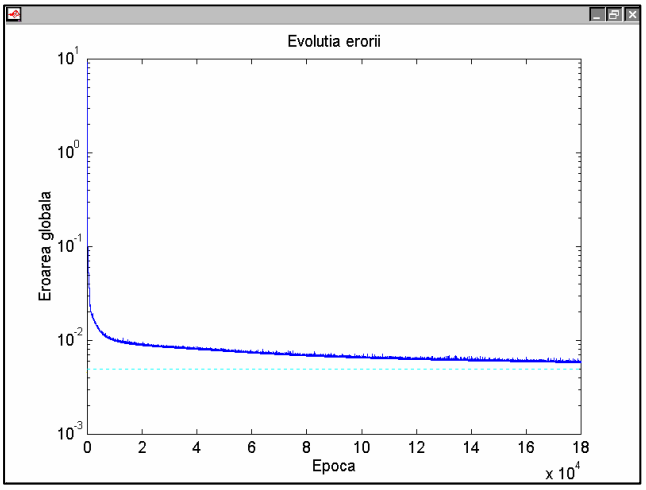

Fig. 7. Evolutia erorii la antrenarea retelei neuronale pentru categoria 3

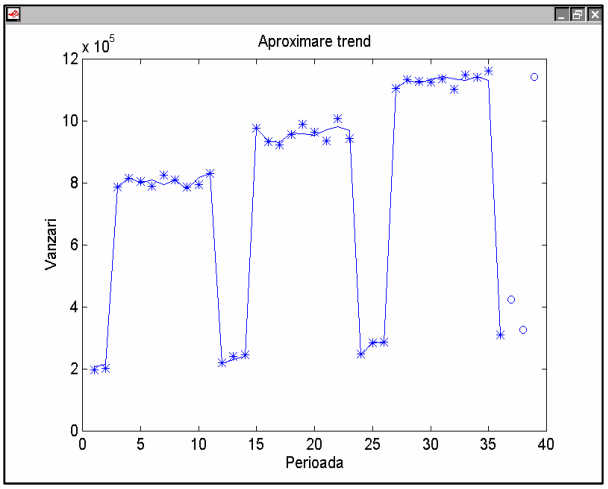

Fig. 8. Aproximarea trendului cu ajutorul retelei neuronale pentru categoria 3

Calculul neuronal prezinta o serie de avantaje, precum: identificarea unor functii de aproximare foarte complexe, acuratete sporita a rezultatelor, o mare varietate de input-uri si output-uri, independenta fata de experti.

Comparând aceste doua metode de AUC în raport cu clasa de cerinte "calitatea rezultatelor si necesarul de resurse" prezentata generic în tabelul 1 am ajuns la urmatoarele concluzii:

1. *Acuratetea rezultatelor*. În analiza multidimensionala acuratetea depinde de abilitatea utilizatorilor de a selecta si utiliza corect modelele predefinite de analiza statistica, pe baza evaluarii vizuale a caracteristicilor de comportament ale faptelor. În multe din cazuri nu exista un model predefinit corespunzator, iar utilizatorii trebuie sa aleaga unul inadecvat. În studiul de caz prezentat, utilizarea unei prognoze exponentiale pentru valoarea vânzarilor genereaza o eroare inacceptabila pentru produsele din categoria 3, în timp ce prognoza cu sezonalitate asigura o eroare acceptabila (tabelul 3). În calculul neuronal, acuratetea rezultatelor depinde de datele de intrare si de calitatea proiectarii retelei neuronale. O retea neuronala bine proiectata poate identifica si aproxima orice caracteristica a datelor de intrare, astfel încât acuratetea va fi foarte mare. În tabelul 3 se prezinta rezultatele valorice, care confirma acuratetea mare pentru calculul neuronal în domeniul vânzarilor fata de analiza multidimensionala.

Tabelul 3: Analiza comparativa a rezultatelor valorice ale prognozei

Categoria de produs	Luna	Valuarea actuala	Express		Matlab
			Prognoza exponentiala	Prognoza cu sezonalitate	Valori prognozate
1	IAN98	2296576	2133153	2279460.38	2296419
	FEB98	2279945	2160568.75	2332578.49	2282065
	MAR98	2322283	2188336.86	2281832.24	2332941
	…		…	…	…
	IAN99	-	2486423.10	2619696.72	2562881
	FEB99	-	2518389.17	2655538.02	2514373
	MAR99	-	2550745.94	2654901.27	2599749
Procentul de eroare			**2.84**	**1.00**	**0.76**
2	IAN98	1603956	1503090.74	1636561.04	1604627
	FEB98	1554450	1521920.48	1566818.46	1548870
	MAR98	1640082	1540986.10	1595496.05	1641169
	…		…	…	…
	IAN99	-	1745284.92	1940963.41	1826476
	FEB99	-	1767148.71	1873342.82	1710930
	MAR99	-	1789286.40	1915427.07	1753852
Procentul de eroare			**3.07**	**1.25**	**0.69**
3	IAN98	284998	764641	282066.16	286905
	FEB98	286231	778461	285667.17	285212
	MAR98	1105907	792530.93	1113696.34	1104408
	…		…	…	…
	IAN99	-	948002.36	335888.51	424112
	FEB99	-	965136.50	340898.56	325655
	MAR99	-	982580.31	1332685.64	1142169
Procentul de eroare			**65.59**	**1.64**	**1.57**

2. *Explicabilitatea proceselor de prelucrare necesare pentru obtinerea rezultatelor.* În analiza multidimensionala majoritatea prelucrarilor tin doar de competenta utilizatorilor, astfel încât explicabilitatea este evidenta. În calculul neuronal explicabilitatea este scazuta din cauza specificitatii schemei de reprezentare a cunostintelor si a gradului mare de integrare a pasilor de prelucare. Pentru a creste gradul de explicabilitate în calculul neuronal, am extras regulile din retelele neuronale, dupa faza de învatare, aplicând câteva metode uzuale ([1], [6]), dar rezultatele au fost neconcludente.

3. *Viteza de raspuns.* În analiza multidimensionala, selectarea modelului de previziune si calcularea valorilor previzionate s-au realizat rapid. În abordarea neuronala, faza de antrenare a retelei a fost de durata

mai mare (tabelul 2), dar apoi valorile previzionate au fost rapid calculate.

4. *Toleranta la zgomote.* Metodele de previziune predefinite în ORACLE Express au tratat foarte bine aparitia zgomotului în date (tabelul 4). În faza de antrenare, o retea neuronala pentru prognoza nu poate identifica zgomotul în datele reale.

Tabelul 4 - Tratarea zgomotului în prognoza valorilor

Luna	Valoarea actuala	Prognoza exponentiala	Prognoza cu sezonalitate
NOE96	1,708,483	1,913,989.51	1,712,180.14
DEC96	1,759,802	1,933,564.86	1,747,569.26
IAN96 (zgomot)	11,831,746	1,953,340.42	4,435,909.49
FEB96	1,969,470	1,973,318.23	4,868,775.84
MAR96	1,975,974	1,993,500.36	2,419,024.36
APR96	2,008,928	2,013,888.90	2,020,737.69
Procentul de eroare		**8.06%**	**16.69%**

5. *Toleranta la complexitate.* Modelele de analiza implementate în ORACLE Express pot manipula datele cu trend si sezonalitate. Pentru dinamici mai complexe de evolutie aceste metode sunt improprii. Abordarea neuronala poate trata orice tip de dinamica.

În concluzie, rezultatele obtinute în studiul de caz confirma aprecierile generale facute în analiza comparativa a principalelor metode de AUC.

Bibliografie

[1] Bodea C. – *Inteligenta artificiala si sisteme expert*, Editura Inforec, Bucuresti, 1998.

[2] Brand E, Gerritsen R. – *Data Mining and Knowledge Discovery*, in: DBMS, Data Mining Solutions Supplement, November, 1998, http://www.dbmsmag.com/9807m01.html.

[3] Dhar V., Stein R. – *Seven Methods for Transforming Corporate Data into Business Intelligence*, Prentice Hall, 1997.

[4] Fayyad U.M., Piatetsky-Shapiro, Smyth P., Uthurusamy R. – *Advances in Knowledge Discovery and Data Mining*, AAAI Series, 1998.

[5] Information Discovery Inc. – *Perspective on Data Mining*, http://www.datamining.com/datamine/dm-ka.htm

[6] Kasabov N.(ed) – *Brain-like Computing and Intelligent Information Systems*, Springer-Verlag, 1998.

[7] Munakata T. – *Fundamentals of the New Artificial Intelligence*, Springer-Verlag, 1998

[8] *** - *MATLAB. High Performance Numeric Computation and Visualisation Software*, Reference Guide, The MathWorks Inc., Natick, Massachusetts, 1992-1998.

[9] *** - *ORACLE Data Warehousing*, SAMS Publishing, 1997

[10] *** - *Express Analyzer*, Oracle Publishing, 1998