TEHNICI DE CLASIFICARE AUTOMATĂ A DATELOR

Utilizarea reţelelor neuronale în stabilirea ratingului de ţară

Această lucrare este rezultatul proiectului „Doctorat şi doctoranzi în triunghiul educaţie – cercetare - inovare (DOC-ECI)", POSDRU/6/1.5/S/11, proiect cofinanţat din Fondul Social European prin Programul Operaţional Sectorial Dezvoltarea Resurselor Umane 2007-2013 şi coordonat de Academia de Studii Economice din Bucureşti

ANA MARIA MIHAELA IORDACHE

TEHNICI DE CLASIFICARE AUTOMATĂ A DATELOR

Utilizarea reţelelor neuronale în stabilirea ratingului de ţară

EDITURA UNIVERSITARĂ
Bucureşti

Colecţia ŞTIINŢE ECONOMICE

Referenţi ştiinţifici: Prof. univ. dr. Elena-Carmen Pintilescu, Universitatea "Alexandru Ioan Cuza" din Iași, Facultatea de Economie şi Administrarea Afacerilor
Prof. univ. dr. Spircu Liliana, Academia de Studii Economice din Bucureşti, Catedra de Cibernetică Economică
Prof. Univ. Dr. Adrian Victor Bădescu, Academia de Studii Economice din Bucureşti, Departamentul de Informatică și Cibernetică Economică
Prof. univ. dr. Aivaz Kamer-Ainur, Universitatea „Ovidius" Constanţa, Facultatea de Ştiinţe Economice

Redactor: Gheorghe Iovan
Tehnoredactor: Ameluţa Vişan
Coperta: Monica Balaban

Editură recunoscută de Consiliul Naţional al Cercetării Ştiinţifice (C.N.C.S.) şi inclusă de Consiliul Naţional de Atestare a Titlurilor, Diplomelor şi Certificatelor Universitare (C.N.A.T.D.C.U.) în categoria editurilor de prestigiu recunoscut.

Descrierea CIP a Bibliotecii Naţionale a României
IORDACHE, ANA MARIA MIHAELA
Tehnici de clasificare automată a datelor : utilizarea reţelelor neuronale în stabilirea ratingului de ţară / Ana Maria Mihaela Iordache. - Bucureşti : Editura Universitară, 2022
Conţine bibliografie
ISBN 978-606-28-1435-9

004

DOI: (Digital Object Identifier): 10.5682/9786062814359

Editura Universitară
Editor: Vasile Muscalu
B-dul. N. Bălcescu nr. 27-33, Sector 1, Bucureşti
Tel.: 021.315.32.47
www.editurauniversitara.ro
e-mail: redactia@editurauniversitara.ro

Distribuţie: tel.: 021.315.32.47/ 0745 200 718/ 0745 200 357
comenzi@editurauniversitara.ro
www.editurauniversitara.ro

CUPRINS

INTRODUCERE

O dată cu apariţia calculatoarelor şi cu mărirea volumului de date a apărut necesitatea de a identifica cunoştinţe noi, necunoscute până în acel moment într-un timp relativ scurt.

În acest context, *data mining*, cunoscută şi ca „descoperirea cunoştinţelor în baze de date mari" este un instrument modern şi puternic al tehnologiei informaţiei şi comunicaţiilor, instrument ce poate fi folosit pentru a extrage informaţii utile dar încă necunoscute.

Acest lucru automatizează procesul de descoperire al unor relaţii şi combinaţii în datele brute, iar rezultatele găsite ar putea fi încadrate într-un sistem automat de suport al deciziei.

Orice activitate umană, atât la nivel microeconomic, cât şi la nivel macroeconomic se desfăşoară în condiţii de risc şi incertitudine. Unele dintre acestea pot fi evitate mai uşor sau mai greu în funcţie de nivelul de cunoaştere, de gradul de evaluare sau de importanţa care li se acordă în fundamentarea deciziilor.

Cercetarea a avut ca obiective principale: prezentarea cât mai exhaustivă a tehnicilor de clasificare automată a datelor având ca suport de decizie reţelele neuronale şi elaborarea unei metodologii de evaluare a ratingului de ţară în funcţie de mai mulţi indicatori aleşi din diverse grupe: indicatori economici, indicatori financiari şi indicatori sociali. Modelul de clasificare propus are aplicabilitate atât la nivel macroeconomic, cât şi la nivel microeconomic.

La nivel microeconomic, în contextul globalizării, derularea de afaceri de către agenţii economici pe pieţele externe se va realiza numai dacă există un stimulent suficient de puternic, în măsură să motiveze companiile să-şi asume riscurile implicate desfăşurării activităţii într-o anumită ţară. Ratingul de ţară calculat de diverse instituţii este un indicator agregat foarte important atunci când se analizează oportunitatea de a investi sau nu în acea zonă.

La nivel macroeconomic, analiza riscului de ţară presupune identificarea dificultăţilor care pot să apară în onorarea de către un anumit stat a obligaţiilor care decurg din angajamentele sale luate pe plan extern.

Lucrarea de doctorat este alcătuită din cinci capitole. În primul capitol intitulat „Aspecte generale privind data mining şi modelarea pe baza reţelelor neuronale" am prezentat un istoric şi o definiţie a termenului data mining, o scurtă sinteză a metodelor şi tehnicilor data mining, punând accent pe modelarea bazată pe reţelele neuronale.

Dezvoltarea calculatoarelor şi implicit a puterii de calcul precum şi a volumului mare de date au condus la necesitatea de a descoperi informaţii noi într-un timp relativ scurt. În acest context, începând cu anii 2000 s-a dezvoltat o nouă tehnologie denumită *data mining*. Metodele *data mining* provin din statistică, administrarea bazelor de date şi din inteligenţa artificială (reţele neuronale, analiza datelor, procesarea imaginilor, învăţarea asistată de calculator, algoritmi genetici, etc.). Printre tehnicile de *data mining* cele mai des întâlnite se numără: excluderea, clasificarea (clusterizarea), discriminarea şi previziunea.

Reţelele neuronale sunt utilizate pentru soluţionarea unor probleme ce nu pot fi rezolvate cu ajutorul algoritmilor convenţionale, cum ar fi probleme de optimizare, probleme de clasificare, etc. O reţea neuronală este alcătuită din neuroni, grupaţi în straturi, fiecare

neuron dintr-un anumit strat fiind conectat la toţi neuronii din stratul anterior şi la toţi neuronii din stratul ce urmează, cu excepţia straturilor de intrare şi ieşire. Cel mai important şi cel mai des folosit algoritm de antrenare a unei reţele neuronale este algoritmul back-propagation.

În capitolul doi al lucrării intitulat „Problema clasificării obiectelor în abordarea ratingului de ţară" am realizat o trecere în revista a stadiului actual al cercetării în domeniul clasificării obiectelor şi am studiat reţeaua neuronală în problema de clasificare.

Reţelele neuronale sunt flexibile, se pot adapta cu uşurinţă datelor de antrenament şi pot gestiona mai bine „zgomotul" şi schimbările apărute în exemplele de învăţare. Odată încheiat procesul de instruire (antrenare), o reţea neuronală poate genera un *output* rezonabil pentru datele de intrare care nu i-au mai fost prezentate în faza de învăţare.

În domeniul clasificării obiectelor cel mai utilizat model este perceptronul multistrat. Acest model presupune existenţa unei reţele neuronale conectată în totalitate, având neuronii dispuşi pe mai multe straturi astfel: strat de intrare, unul sau mai multe straturi ascunse şi un strat de ieşire. Printre avantajele perceptronului multistrat se numără: eficienţa în reamintire şi eficienţă mare în clasificare. Timpul îndelungat de instruire al acestui tip de reţea neuronală limitează oarecum aplicarea modelului. Mai departe am prezentat pe scurt metodologia folosită de diverse agenţii pentru evaluarea ratingului de ţară şi a modului de acordare al calificativului pentru un anumit stat.

Pornind de la definiţia conform căreia riscul reprezintă variabilitatea rezultatului posibil în funcţie de un eveniment nesigur, incert, există mai multe agenţii de rating, fiecare dintre acestea având propriul model de evaluare al riscului şi oferind diverse produse pentru diferite segmente de utilizatori. Astfel, principalele agenţii de rating existente sunt: Institutional Investor, Standard & Poor`s, Political Risk Services, The Economist, Moody`s şi altele.

La realizarea aplicaţiei am utilizat următoarea metodologie: alegerea grupului de ţări asupra căruia se realizează aplicaţia (ţările din Uniunea Europeană: Austria, Danemarca, Suedia, Finlanda, Germania, Cipru, Letonia, Polonia, Lituania, Olanda, Spania, Portugalia, Franţa, Italia, România, Grecia, Marea Britanie, Bulgaria, Belgia, Ungaria, Malta, Slovacia, Cehia, Estonia, Slovenia, Luxemburg, Irlanda); stabilirea indicatorilor, înregistrarea valorilor acestora (pentru anul 2009) şi aplicarea tehnicilor de analiză a datelor (analiza în componente principale, analiza cluster şi analiza factorială) asupra fiecărei grupă de indicatori; extragerea indicatorilor care aduc aport informaţional maxim din fiecare grupă; realizarea şi instruirea reţelei neuronale; interpretarea rezultatelor obţinute şi realizarea de previziuni (simulări) asupra situaţiei României.

Partea a treia a lucrării prezintă aplicabilitatea tehnicilor de *data mining*, mai precis extragerea informaţiilor "ascunse", "importante" din listele cu indicatorii ţărilor înregistraţi în diferite planuri (economic, social şi financiar) folosind astfel tehnici de analiză a datelor (analiză în componente principale, analiză factorială şi analiză cluster).

Seturile de indicatori utilizaţi în aplicaţie sunt următorii:

- Indicatori economici: consum final (% din PIB), cererea internă (% din PIB), consumul final al gospodăriilor (% din PIB), formarea brută de capital (% din PIB), formarea brută de capital fix (% din PIB), exporturi (% din PIB), importuri (% din PIB), valoarea adăugată brută (preţuri curente) (% din PIB), impozite fără subvenţii pe produs (% din PIB), salarii (% din PIB), formarea brută de capital fix de către sectorul privat (% din PIB);

- Indicatori sociali: pondere angajaţi cu contract parţial de muncă în total angajaţi, ponderea populaţiei angajate în total populaţie, ponderea populaţiei în vârstă 0-14 ani în total populaţie, ponderea populaţiei în vârstă 15-24 ani în total populaţie, ponderea populaţiei în vârstă 25-49 ani în total populaţie, ponderea populaţiei în vârstă 50-64 în total populaţie,

ponderea populaţiei în vârstă 65-79 în total populaţie, ponderea populaţiei în vârstă de peste 80 ani în total populaţie, rata şomajului, vârsta medie a populaţiei;

- Indicatori financiari: rata anuală medie a inflaţiei (raportată la indicele preţurilor de consum), deficit public fiscal (% din PIB), investiţii publice fixe (% din PIB), datoria publică (guvernamentală) % din PIB, soldul contului curent (% PIB), împrumuturi guvernamentale externe (% din PIB).

Pentru fiecare set de indicatori am aplicat tehnici de analiză a componentelor principale, tehnici de analiză cluster şi analiză factorială pentru o primă clasificare a ţărilor în funcţie de indicatorii din fiecare grupă.

Ultimul capitol al lucrării intitulat ”Clasificarea ţărilor prin intermediul reţelei neuronale” prezintă asocierea unei reţele neuronale pentru problema de clasificare a ţărilor în funcţie de indicatorii care aduc maximum de informaţie din cele trei grupe de indicatori analizate în capitolul patru.

Pentru a asocia o reţea neuronală de tip perceptron multistrat problemei analizate este necesar să se cunoască mai întâi care sunt indicatorii care aduc maximum de informaţie din fiecare grupă analizată şi apoi în câte clase se împart ţările în funcţie de aceşti indicatori.

În urma aplicării tehnicilor de analiză a datelor indicatorii care aduc aport informaţional maxim din fiecare grupă sunt: cererea internă, formarea brută de capital, exporturi (din domeniul economiei reale); ponderea populației angajate în total populație, ponderea populației în vârstă de 15-24 ani în total populație, vârsta medie a populației (din plan social); deficit public fiscal, investiții publice fixe, împrumuturi guvernamentale externe (din domeniul financiar).

După instruirea reţelei neuronale s-a constatat că sunt ţări care şi-au schimbat clasa în care erau iniţial. Acest lucru a fost posibil datorită faptului că statele din Uniunea Europeană au fost împărţite în trei părţi: date de antrenare (80% din totalul ţărilor), date de validare şi date de testare.

Tot în acest capitol se prezintă anumite previziuni (scenarii) de îmbunătăţire a performanţelor economice, sociale şi financiare ale României cu scopul de a accede în clasa ţărilor dezvoltate. Previziunile au fost obţinute pe baza scorurilor şi a funcţiilor de transfer prin reţeaua neuronală deja antrenată.

La nivelul Uniunii Europene este importantă stabilirea de politici economice, sociale și/sau financiare cu scopul evitării dezechilibrelor macroeconomice și urmărirea realizării unei creșteri sustenabile sprijinită, acolo unde este necesar, de adoptarea unor reforme structurale. Pentru realizarea acestui lucru, în ultimii ani se vorbește tot mai mult de un nou concept numit macroprudențialitate. Dar oricât de eficiente ar fi politicile macroprudențiale, acestea nu pot înlocui politicile macroeconomice, sociale sau financiare adoptate la nivelul fiecărui stat.

CAPITOLUL 1

ASPECTE GENERALE PRIVIND *DATA MINING* ȘI MODELAREA PE BAZA REȚELELOR NEURONALE

Termenul *data mining* presupune analiza datelor din diverse puncte de vedere (aspecte) cu scopul de a extrage cunoștințe pentru a le folosi mai departe în fundamentarea deciziilor la nivel micro sau macroeconomic. Din cele mai vechi timpuri oamenii au fost atrași de funcționarea creierului uman și de modul de adoptare a deciziilor de către acesta. În acest sens au început să studieze rețelele neuronale, să elaboreze diverși algoritmi de instruire a acestora. Datorită volumului mare de informație rezultat pe baza modelării neuronale au fost dezvoltate și rețelele hebbiene, principala lor caracteristică fiind introducerea în calcul a unui factor de uitare cu scopul de a păstra mai departe informația cu adevărat importantă.

1.1 Istoric şi definirea termenului de *data mining*

Data mining s-a dezvoltat ca o consecinţă a dezvoltării bazelor de date. Colectarea datelor în diverse formate electronice a început în anii '60 permiţând o analiză retrospectivă a datelor prin intermediul calculatorului. Bazele de date relaţionale au apărut în anii '80 împreună cu Structured Query Language (SQL). Anii '90 sunt caracterizaţi de o explozie a informaţiilor. Pentru stocarea lor au început să se folosească depozitele de date (data warehouses). ***Data mining*** a apărut ca răspuns la provocările cu care s-au confruntat comunitatea specialiştilor în baze de date, care se ocupau cu cantităţi masive de date, aplicarea analizei statistice şi aplicarea tehnicilor de căutare, specifice inteligenţei artificiale, asupra datelor.[1]

Principalele domenii care pot fi considerate surse de date sunt:

- mediul economic;
- telecomunicaţii (în telefonie se generează anual aprope 20 exabytes de date);
- sateliţi; internet – ordin de mărime: terabytes;
- biblioteci: (Library of Congress 20tb-3pb date);
- agenţii legale (baza de date cu amprentele FBI: 1 petabyte);
- experimente ştiinţifice.

Data mining este aplicat într-o varietate de domenii, începând cu managementul de investiţii până la astronomie. Importanţa şi potenţialul de aplicare al ***data mining*** a fost recunoscut în marketing, domeniul bancar, asigurarea sănătăţii, telecomunicaţii, pentru aplicaţii cum ar fi analiza coşului de piaţă, pentru promovarea eficienţei, analiza vulnerabilităţii clienţilor, managementul relaţiilor cu clienţii, crearea de portofoliu, detectarea fraudei în telefonia celulară etc.

În 1995, Gartner Group Advanced Technology Research Note a poziţionat *data mining* şi inteligenţa artificială pe primul loc între cele cinci zone tehnologice cheie care "vor

[1] Ionita Angela, Asupra termenului de minerit de date (data mining), http://rria.ici.ro/ria2005_2/art03.html

avea în mod clar impact peste un spectru larg de domenii industriale în următorii trei până la cinci ani".[2]

Data mining este un pas în procesul KDD (knoweldge discovery in data – descoperirea cunoştinţelor din date), care constă în aplicarea de algoritmi de descoperire şi analiză de date, care sub limitele eficienţei de calcul vor produce o mulţime de modele pe datele studiate (figura 1).

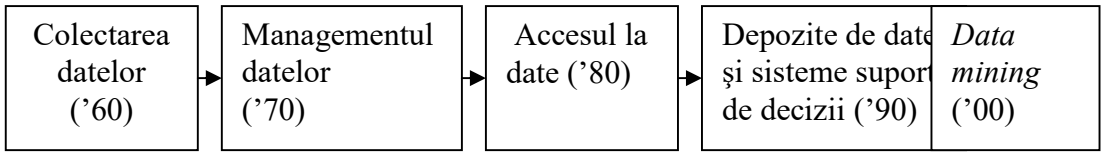

Figura 1. Evoluţia tehnologiei spre *data mining*

Data mining reprezintă o arie de intersecţie între învăţarea asistată de calculator, statistică şi baze de date.[3]

Data mining reprezintă procesul de selecţie, explorare şi modelare a unor seturi mari de date pentru a descoperi modele necunoscute, utile în dezvoltarea afacerilor.[4]

Fayyad U., Piatetsky-Shapiro G., Smyth P afirmau că „***data mining*** este procesul de identificare de cunoştinţe valide, noi, potenţial utile şi, în final, inteligibile din baze de date care sunt folosite în luarea deciziilor hotărâtoare în domeniul afacerilor, mai precis, „***data mining*** reprezintă aplicarea unor algoritmi specifici pentru extragerea modelelor din date”. [5]

Din punct de vedere al dicţionarului explicativ al limbii romane, ***data mining*** este o expresie improprie care asociază procesul de căutare şi găsire a cunoştinţelor dintr-o cantitate mare de date cu procesul de extragere a mineralelor din roci. Ambele procese se refera la extragerea esenţialului din domenii diferite.

În conformitate cu Carta Alba a soluţiilor de management de date de la IBM**,** ***data mining*** este procesul de extragere de informaţii valide, necunoscute anterior şi, în final, inteligibile din baze mari de date, informaţii folosite în luarea deciziilor hotărâtoare în domeniul afacerilor. Extragerea informaţiilor poate fi utilizată la formarea de modele de predicţie sau de clasificare, pentru a identifica relaţii între înregistrările din bazele de date sau pentru a furniza un sumar al bazelor de date care sunt minerite. ***Data mining*** constă dintr-un număr de operaţii, fiecare fiind suportată de o varietate de tehnici cum sunt reguli de inducţie, reţele neuronale, clusterizare conceptuală, descoperire asociativă etc. În multe domenii din lumea reală cum ar fi analiza de marketing, analiza financiară, detectarea fraudei, informaţiile extrase necesită utilizarea cooperativă a mai multor operaţii şi tehnici de minerit de date.[6]

Toate cele trei aspecte provin din tehnologiile de *data mining*, specifice diferitelor domenii, şi se ocupă de problema uşurinţei în utilizare prin ascunderea complexităţii algoritmilor fundamentali de *data mining*.

„...Pentru a concluziona, instrumentele de *data mining* sunt orientate către a obţine abilităţi de analiză multimedia şi capacităţi de analizare simultană a numeroase tipuri de baze de date. Utilizatorii finali sau specialiştii în *data mining* vor fi persoane care operează cu

[2] http://www.thearling.com/text/dmwhite/dmwhite.htm
[3] *Marcel Holsheimer: Data Mining by Business Users: Integrating Data Mining in Business Processes. KDD Tutorial Notes, 1999*
[4] *SAS Institute*
[5] *Fayyad U.M., Piatetsky-Shapiro G., Smyth P.,* ***Knowledge Discovery in Database,*** *AAAI Press, 1996*
[6] IBM's Data Mining Technology - White Paper data Management Solutions, April, 1996

instrumentele care sunt înglobate în pachetele software standard, cel mai probabil utilizând tehnologii de reţele neuronale."[7]

Alte definiţii ar putea fi:

- „***data mining*** este un proces de descoperire de relaţii, şabloane şi cunoştinţe din date;
- „***data mining*** este procesul prin care informaţiile şi cunoştinţele sunt extrase din volumele mari de date folosind tehnici care sunt mai mult decât o simplă căutare în date".

Principalele discipline care se intersectează cu domeniul *data mining* sunt: statistica, programarea calculatoarelor, învăţarea asistată de calculator, tehnologia bazelor de date, tehnologia digitală, tehnologia informaţiei, etc (figura 2).

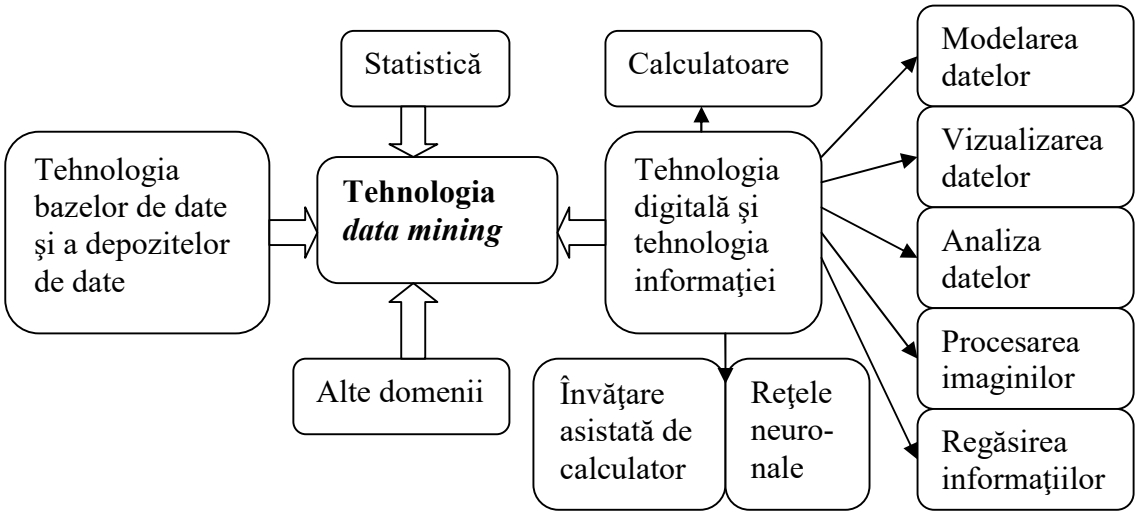

Figura 2. Discipline cu care se intersectează *data mining*

În concluzie, *data mining* a evoluat ca o tehnologie datorită a două fenomene complementare:

- volumul de date în continuă expansiune în urma dezvoltării tehnologiilor de baze de date şi a instrumentelor de colecţionare a datelor;

- nevoia de cunoaştere concretizată prin necesitatea de a filtra şi interpreta toate aceste volume de date stocate în baze de date, depozite de date sau bănci de date.

Procesul de descoperire a cunoştinţelor din cantităţi mari de date este un proces complex format, în principal, din următoarele **etape**, care se succed din punct de vedere cronologic:

- *curăţirea datelor* (data cleaning) se realizează prin înlăturarea datelor inutile şi a celor inconsistente;

- *integrarea datelor* (data integration) constă în combinarea datelor provenite din mai multe surse de date diferite;

- *selectarea datelor* (data selection) reprezintă extragerea datelor relevante pentru analiză din depozitele de date disponibile (baze de date, depozite de date, www etc.);

- *transformarea datelor* (data transformation) constă în punerea datelor în formate unitare, corespunzătoare pentru *data mining* (analiza în vederea descoperirii de cunoştinţe) prin realizarea unor operaţii rezumative şi/sau de agregare (de unificare);

- *data mining* este definită prin extragerea unor modele de date aplicând asupra datelor rezultate după parcurgerea etapelor anterioare și metode inteligente, denumite generic metode

[7] ATEL, L., S. PERRY, D. TAYLOR, A. BROWN, S. TIKE: The Future of Data Mining

data mining. Acest proces este esenţial pentru descoperirea cunoștinţelor utile „ascunse” în depozitele de date;

- *evaluarea modelelor* (pattern evaluation) se realizează prin evaluarea modelelor de date extrase pentru identificarea celor care reprezintă cunoștinţele care interesează în mod real;

- *prezentarea cunoștinţelor* (knowledge presentation obţinute utilizatorilor lor, prin folosirea unor tehnici de vizualizare şi de reprezentare adecvate.

Metodele *data mining* provin din calculul statistic clasic, din administrarea bazelor de date şi din inteligenţa artificială. Ele nu înlocuiesc metodele tradiţionale ale statisticii, ci sunt considerate a fi extinderi ale tehnicilor grafice şi statistice. Deoarece softului îi lipseşte intuiţia umană, rezultatele metodelor *data mining* vor trebui supuse în mod sistematic unei supravegheri umane.

Structura tipică de date potrivită pentru *data mining* conţine observaţiile plasate pe linii iar variabilele plasate pe coloane. Domeniile sau intervalele de valori pentru fiecare variabilă vor trebui sa fie definite precis, evitându-se cât mai mult posibil exprimările vagi.

1.2 Metode şi tehnici de *data mining*

Rezultatele obţinute cu ajutorul tehnicilor de *data mining* pot varia şi sunt specifice fiecărui tip de utilizator. În general, tehnicile de *data mining* se aplică într-o bază de date din două motive:

- validarea unei ipoteze privind datele, corectitudinea acestora, dar şi ipoteze statistice;

- descoperirea de noi caracteristici din date. Descoperirea, la rândul ei, poate fi împărţită în descriere si predicţie. Descrierea datelor se realizează, fie prin calcularea de diverşi indicatori statistici (elementari) din datele brute, cum ar fi media, dispersia, abaterea medie pătratică, etc., fie prin aplicarea asupra datelor a tehnicilor avansate de analiză, cum ar fi analiza cluster, analiza componentelor principale, analiza discriminantă, etc. Predicţia are ca scop exprimarea cât mai exactă a unor valori viitoare pentru datele aflate în analiză.

Cele mai folosite tehnici de *data mining* sunt:

1. **Excluderea** ce presupune tratarea din punct de vedere informatic a datelor. Aceasta presupune identificarea şi eliminarea înregistrărilor care conţin anomalii în comparaţie cu celelalte date (date aberante, înregistrări cu date lipsa, etc).

2. **Clasificarea (clusterizarea)** este operaţia prin care obiectele dintr-o mulţime dată sunt repartizate în submulţimi numite „clase” în funcţie de asemănările şi deosebirile dintre ele.

3. **Discriminarea** este diferită de clasificare din cauza necesităţii, în aplicaţii, a unor cunoştinţe anterioare asociate claselor. De obicei este necesar un eşantion din fiecare clasă pentru a realiza o analiză discriminantă. Analiza discriminantă are următoarele obiective: investigarea diferenţelor între grupuri, separarea eficientă a grupurilor, identificarea variabilelor importante discriminatorii, testarea ipotezelor legate de diferenţele dintre grupuri, clasificarea de noi observaţii în grupuri pre-existente.

4. **Previzionarea** ce se realizează pe baza trendului.

Daca ne referim la metodele „de învăţare” (*data mining*) acestea pot fi grupate în doua categorii: nesupervizate, respectiv supervizate.

Metodele de învăţare nesupervizate includ următoarele:

1) Analiza componentelor principale (ACP). Scopul său este de a reduce dimensionalitatea datelor multivariate prin „integrarea” variabilelor corelate, transformând liniar variabilele iniţiale în variabile necorelate între ele.

2) Analiza factorială (Factor Analysis) permite extragerea unui număr mic de factori ascunşi care explică cea mai mare parte a variabilităţii comune şi determină corelaţiile observate între datele iniţiale.

3) Analiza clasificării (Cluster Analysis) are ca scop de a grupa cazurile (observaţiile) în clustere (grupuri, categorii).

Printre metodele de învăţare supervizate se numără:

1) Regresia liniară multiplă având scopul de a descrie variaţiile unei variabile *y* în funcţie de valorile altor variabile prin intermediul unor funcţii liniare sau liniarizabile. Variabila *y* se numeşte dependentă, iar celelalte poartă denumirea de variabile independente sau explicative.

2) Regresia logistică este un caz particular al regresiei liniare multiple. În acest tip de regresie „răspunsul" este o variabilă binară de tipul „da sau nu".

3) Analiza discriminării este o tehnica de statistică multivariată frecvent utilizată pentru a construi un model descriptiv/predictiv al separării („discriminării") pe grupuri bazat pe variabilele predictor observate şi de a clasifica fiecare observaţie într-unul din grupuri. În analiza discriminantă sunt utilizate mai multe atribute cantitative pentru a separa o singură variabilă de clasificare.

4) Reţelele neuronale asociază unui input vectorial un anumit output vectorial. Reţelele neuronale oferă modele matematice de grupare neliniare în special pentru clasificarea unor obiecte având la bază conceptul de reţea bioelectrică din creierul uman formată de neuroni şi sinapsele acestora. Principala trăsătură a acestor reţele este capacitatea de a învăţa pe bază de exemple, folosindu-se de experienţa anterioară pentru a-şi îmbunătăţi performanţele.

În scopul obţinerii beneficii maxime din aplicarea algoritmilor de *data mining* asupra unui set de date este necesară asigurarea integrităţii şi valabilităţii datelor. Principalele acţiuni care se au în vedere sunt[8]: curăţirea, transformarea, sintetizarea, reducerea si discretizarea datelor. (Kennedy 1998, Pzle 1999)

Curăţirea datelor. Datele din lumea reală sunt de obicei incomplete şi afectate de zgomot. Curăţirea lor presupune adăugarea valorilor lipsă, reducerea zgomotului şi înlăturarea valorilor aberante. Valorile lipsă provin cel mai adesea din: disfuncţionalităţi ale echipamentelor de colectare, neconcordanţe între datele într-un set de date sau omisiunea acestora, atunci când datele nu sunt înţelese sau sunt considerate banale. Practicile cele mai folosite pentru tratarea acestor cazuri sunt: ignorarea întregului tuplu din baza de date, completarea manuală a datelor lipsă, folosirea unui caracter special care arată o valoare lipsă, folosirea unor aproximări ale valorilor lipsă.

Sintetizarea datelor presupune efectuarea unei sinteze exhaustive a bazelor de date multidimensionale, băncilor de date sau a fişierelor într-o schemă unitară. Sintetizarea datelor se poate realiza pe trei niveluri (Han si Kamber, 2001): sintetizarea schemei bazei de date, detectarea şi rezolvarea neconcordanţelor dintre valorile datelor (de exemplu să se păstreze aceeaşi scală a unităţii de măsură), managementul datelor redundante.

Transformarea datelor. Cele mai utilizate tehnici de transformare a datelor sunt: eliminarea zgomotului din date, agregarea, generalizarea, construirea de caracteristici (atribute) noi şi normalizarea. Cele mai utilizate tehnici de normalizare, potrivit cercetătorilor Weiss şi Indurkhya sunt: (Weiss and Indurkhya, 1998)

[8] Negnevitsky Michael, *Artificial Intelligence – A guide to intelligent systems, second edition*, Editura Pearson Education, England, 2005, ISBN 0321-20466-2

- normalizarea min-max presupune aplicarea unei transformări liniare asupra datelor:

$$x' = \frac{x - min}{max - min}, x' \epsilon [0, 1].$$

- normalizare după scorul z. Un atribut A este normalizat în funcţie de medie şi dispersie $x' = \frac{x - media_x}{dispersia_x}$.

- normalizare prin scalare zecimală se realizează prin aplicarea formulei: $x' = \frac{x}{10^j}$, unde *j* este cel mai mic întreg care satisface relaţia $max(|x'|)<1$.

Şi să revenim asupra metodei de **clasificare.** Aceasta reprezintă procesul de găsire a proprietăţilor comune dintr-un set de date şi clasificarea acestora în diferite clase conform unui model de clasificare. (Chen, 1996). Clasificarea presupune parcurgerea a două etape:

- existenţa unui set de învăţare sau de antrenare folosit pentru a construi un model care clasifică datele în anumite grupuri (clase);

- folosirea modelului de clasificare construit pentru un nou set de date sau pentru a extrage reguli de clasificare noi.

Cele mai utilizate tehnici de clasificare sunt: clasificarea bayesiană, arbori de decizie şi reţele neuronale.

Clasificarea Bayesiană presupune asignarea unui element *x* la una din clasele $c_1, c_2, ... c_N$, date, folosind un model de probabilitate definit conform teoremei lui Bayes.

O reţea bayesiană este un graf orientat ale cărui noduri reprezintă variabilele şi ale cărui muchii sunt legăturile cauzale sau de influenţă dintre acestea. Fiecărui nod i se asociază un tabel de probabilităţi ce descriu relaţiile dintre acesta şi părintele său.

Strategiile de decizie Bayes se aplică în probleme de clasificare cu număr mare de clase cu scopul a minimiza erorile, având la bază probabilitatea condiţionată. Probabilitatea condiţionată este folosită pentru a măsura încrederea ca un eveniment aleator să aibă loc ştiind că alt eveniment aleator a avut loc. Fie două evenimente *A* si *B*, probabilitatea condiţionată ca evenimentul *A* să aibă loc ştiind că evenimentul *B* a avut loc este $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$. Fie $(\Omega, \sum, P)$ un spaţiu de probabilitate, *B* un eveniment arbitrar din $\sum$ si $\{A_1, ... , A_n\}$ o partiţie a spaţiului Ω. Atunci, $P(A_i \mid B) = \frac{P(B \mid A_i) \cdot P(A_i)}{P(B)} = \frac{P(B \mid A_i) \cdot P(A_i)}{\sum_{i=1}^{n} P(B \mid A_i) \cdot P(A_i)}$, unde $P(B)>0$ si $P(A_i)>0$, $i=1,...n$. $P(A_i|B)$ se numeşte probabilitate posterioară, $P(A_i)$ este probabilitatea apriorică, $P(B|A_i)$ poartă denumirea de verosimilitate. Scopul este de a minimiza probabilitatea de a greşi.

Fie D_i regula de decizie referitoare la clasa Ω_i. Fiind dat un vector *X*, eroarea relativă la clasa Ω_i este definită de $P\{eroare/x\}=q-P(\Omega_i|x)$. Se minimizează probabilitatea de a greşi. Conform regulii bayesiene de decizie se alege D_j dacă $P(\Omega_j \mid x) > P(\Omega_i \mid x), i \in \{1,..., j-1, j+1,...r\}$.

Pentru un nou set de date care urmează a fi clasificate utilizând un clasificator bayesian, presupunem că fiecare atribut este o variabilă aleatoare. Fiind dat un obiect cu atributele $\{A_1, A_2, ..., A_p\}$ se doreşte asignarea sa în clasa Ω_i. Clasificarea este corectă atunci când probabilitatea condiţionată $P(\Omega_i|A_1, A_2, ..., A_p)$ este maximă. Se calculează probabilităţile posterioare $P(\Omega_i|A_1, A_2, ..., A_p)$ pentru toate clasele Ω_i. Se alege apoi clasa care are probabilitatea maximă.

Arborele de decizie este un arbore care conţine în noduri câte un test pentru o anumită proprietate, fiecare arc având o valoare a proprietăţii testate în nodul din care pleacă arcul respectiv, iar în fiecare nod terminal o clasă. Arborii de decizie sunt folosiţi pentru a selecta cea mai buna direcţie de acţiune în situaţiile în care apare incertitudinea. Nodurile frunză specifică decizia la care se ajunge pentru calea care duce la acel nod.

ID3 (**Iterative Dichotomiser 3)**[9] este un algoritm de construire inductivă a unor arbori de decizie. Euristica algoritmului ID3 măsoară câştigul informaţional pe care îl aduce fiecare atribut şi alege ca test acel atribut care maximizează acest câştig. Pentru o mulţime de mesaje $M = \{m_1, m_2, ..., m_n\}$ şi o probabilitate $p(m_i)$ de apariţie a fiecărui mesaj, conţinutul informaţional al unui mesaj din M se defineşte astfel: $I(M) = \sum_{i=1}^{n} - p(m_i) log_2(p(m_i))$.

Algoritmul ID3 foloseşte teoria informaţiei pentru a selecta atributul care oferă cel mai mare câştig informaţional în clasificarea exemplelor de învăţare. ID3 alege atributul considerat cel mai important, şi împarte exemplele în subseturi corespunzătoare valorilor posibile ale atributului. În continuare se apelează ID3 recursiv pentru subseturile obţinute şi lista de atribute rămase. Procedeul recursiv se termină când toate exemplele dintr-un subset au aceeaşi clasificare. Dacă se termină toate atributele şi subsetul conţine încă exemple de clasificare diferite, atunci înseamnă că sunt exemple cu aceeaşi descriere dar cu clasificare diferită. Acesta se poate datora mai multor motive:

- o parte din date sunt incorecte;
- datele sunt corecte dar atributele sunt insuficiente;
- datele sunt corecte dar clasificarea implică un anumit grad de nedeterminism.

Pentru a rezolva problema valorilor şi a mări viteza de execuţie se poate aplica algoritmul C4.5. Acesta se bazează pe algoritmul ID3, ajutând la îmbunătăţirea adoptării deciziei prin care derivă regulile de clasificare. Algoritmul C4.5 generează un arbore de decizie prin partiţionarea recursivă a mulţimii de date, printr-o strategie de parcurgere în adâncime ("depth-first"). Algoritmul ia în considerare toate testele posibile pentru împarţirea mulţimii de date şi selectează testul care aduce cel mai mare aport (câştig) informaţional. Acurateţea clasificării şi timpului de rezolvare pot fi acceptabile pentru baze de date mici, dar în cazul bazelor de date de dimensiuni mari ori timpul de execuţie creşte, ori acurateţea clasificării descreşte considerabil.

Din punct de vedere informatic clasificarea presupune împarţirea unei banci de date în baze de date mai mici prin algoritmi de tipul „divide-et-impera" (împarte şi domină), pe baza conceptului de similitudine. În procesul de clasificare se poate preciza sau nu de la început numărul de clase.

Pentru ca un proces de clasificare sa fie corect trebuie îndeplinite următoarele condiţii (Han si Kamber, 2001): similaritate între grupuri mare şi similaritate mică în interiorul grupurilor.

Tehnicile de clasificare pot fi împărţite în funcţie de următoarele trei criterii (Jain, 1999):

- în funcţie de metoda identificării clusterilor pot fi clasificate în tehnici de partiţionare ierahice, bazate pe densitate şi pe reţele;
- în funcţie de tipul de date cu care se lucrează, pot fi algoritmi statistici şi conceptuali;
- în funcţie de teoria folosită pentru a extrage clusterii, putem avea clusterizarea fuzzy, clusterizare „clară" şi algoritmi care se bazează pe teoria reţelelor neuronale (Kohonen).[10]

[9] Negnevitsky Michael, Artificial Intelligence – A guide to intelligent systems, first edition, Editura Pearson Education, England, 2002, ISBN 0201-71159-1

[10] Negnevitsky Michael, Artificial Intelligence – A guide to intelligent systems, second edition, Editura Pearson Education, England, 2005, ISBN 0321-20466-2

În continuare voi prezenta pe scurt câtiva algoritmi de clasificare.

Algoritmul *k-means* este un instrument bine cunoscut şi frecvent folosit în clusterizare. Presupunând n vectori de atribute $x_1, x_2, ..., x_n$ şi k clustere, $k < n$, fixat. Fie m_i vectorul valorilor medii a vectorilor din clusterul i. Pentru separarea clusterelor se poate utiliza un clasificator bazat pe distanţa minimă. Se poate spune că un *item* (obiect) x aparţine clusterului i dacă $|| x - m_i ||$ este minimă dintre toate cele k distanţe posibile.

S-a constatat că rezultatele obţinute prin *k-means* depind de alegerea valorilor medii iniţiale ale clusterelor. De aceea, o variantă pentru îmbunătăţirea performanţelor algoritmului este rularea sa multiplă şi alegerea partiţionării celei mai bune. O măsură a calităţii clusterizării este dată de calcularea entropiei claselor rezultate. Din mai multe variante de clusterizare, varianta optimă are cea mai mare sumă a entropiilor, ceea ce arată faptul că datele din clustere sunt cât mai omogene $H_c = -\sum_{i=1}^{k} p_i \log p_i$, $K^* = \arg\max_c H_c$, unde H_c este entropia unei clusterizări, p_i sunt probabilităţile de apartenenţă ale valorilor la un anumit cluster (c), iar K^* este strategia optimă de clusterizare. Printre dezavantajele pe care le are algoritmul *k-means* se numără următoarele: se aplică numai seturilor de date numerice, necesită specificarea numărului de clustere în avans, are capacitate redusă de a lucra cu date aberante sau date care sunt afectate de zgomot, precum şi incapacitatea de a descoperi clustere cu forme neconvexe. Pentru a rezolva aceste inconveniente ale lui *k-means* s-au dezvoltat alţi algoritmi. Algoritmul CLARANS este un algoritm tipic care foloseşte medoizii şi care se bazează pe algoritmii PAM si CLARA. Un medoid este elementul central dintr-o clasă. Acest algoritm se foloseşte pentru a identifica k clase optime. O dată ce s-au identificat cei k medoizi, datele rămase sunt asignate grupului corespunzând celui mai aproape medoid. Apoi se verifică dacă nu se poate înlocui un medoid cu oricare dintre vecini. Dacă orice alt punct îmbunătăţeşte calitatea grupului, atunci acesta devine noul medoid şi procesul se repetă. Dacă nu, se alege un optim local şi se aleg aleator alte date pentru a găsi un optim global. Complexitatea algoritmului CLARANS creşte liniar cu numărul de obiecte.

Pentru a manipula baze de date spaţiale au fost creaţi alti doi algoritmi, care au la bază algoritmul CLARANS: SD (CLARANS) şi NSD (CLARANS), utilizatorul având posibilitatea de a alege tipul de reguli extrase în procesul de învăţare.

Algoritmii ierarhici construiesc arbori în care fiecare nod reprezintă un cluster. Mai departe ei se pot împarţi în algoritmi cumulativi şi algoritmi separatori, urmărind o abordare de jos în sus sau de sus în jos. Cei mai utilizaţi algoritmi sunt:

- CURE: Un algoritm robust care lucrează cu valori aberante şi poate identifica clusteri din forme nesferice (Ghula, 1998);

- ROCK: un algoritm de clusterizare robust folosit pentru date booleene şi categoriale. El introduce două noi concepte: vecinii unui punct şi legăturile lor (Guha, 2000);

- Birch este un algoritm care se utilizează pentru a clasifica seturi mari de date şi are la bază două elemente: caracteristica după care se face gruparea şi arborele de grupare. Acest algoritm construieşte o dendrogramă numită arbore de clasificare în timp ce inspectează setul de date. Există două etape cheie în aplicarea algoritmului: scanarea (parcurgerea) bazei de date şi construirea unui arbore şi apoi aplicarea algoritmului de grupare pentru a clasifica noduri «frunză». [11]

[11] Holsheimer Marcel, Data Mining by Business Users: Integrating Data Mining in Business Processes, San Diego, California, USA, 1999, ISBN 1-58113-171-2

Fiind dat un număr de N obiecte $\{Q_i\}$ se calculează următorii indicatori pentru a determina similaritatea:[12]

- centroid-ul (centrul de greutate) $X_o = \frac{\sum_{i=1}^{N} X_i}{N}$;

- raza (distanţa medie de la puncte la centrul de greutate): $R = \left(\frac{\sum_{i=1}^{N} (X_i - X_0)^2}{N} \right)^{\frac{1}{2}}$;

- diametrul (distanţa medie dintre oricare două elemente X_i si X_j ale unui cluster):

$$D = \left(\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(X_i - X_j)^2}{N(N-1)} \right)^{\frac{1}{2}} ;$$

- măsurarea gradului de asemănare între două obiecte X_i si X_j. În general se utilizează metrica lui Minkowski: $d_p(X_i - X_j) = \left(\sum_{k=1}^{N} \left| X_{ik} - X_{jk} \right|^p \right)^{\frac{1}{p}}$, unde N este dimensiunea obiectului (lungimea vectorului, etc). Distanţa euclidiană se obţine pentru p=2, iar distanţa Manhattan pentru p=1;

- media între două clustere: $D2 = \left(\frac{\sum_{i=1}^{N_1}\sum_{j=N_1+1}^{N_1+N_2} (X_i - X_j)^2}{N_1 N_2} \right)^{\frac{1}{2}}$, unde N_1 si N_2 reprezintă dimensiunea primului cluster, respectiv celui de-al doilea cluster;

- media în interiorul clusterului: $D3 = \left(\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} (X_i - X_j)^2}{N(N-1)} \right)^{\frac{1}{2}}$, unde N reprezintă dimensiunea clusterului;

- creşterea varianţei:

$$\left(\sum_{k=1}^{N_1+N_2} \left(X_k - \frac{\sum_{l=1}^{N_1+N_2} X_l}{N_1+N_2} \right)^2 - \sum_{i=1}^{N_1} \left(X_i - \frac{\sum_{l=1}^{N_1} X_l}{N_1} \right)^2 - \sum_{j=N_1+1}^{N_1+N_2} \left(X_j - \frac{\sum_{l=N_1+1}^{N_1+N_2} X_l}{N_2} \right)^2 \right)^{\frac{1}{2}} .$$

[12] Anghelache Constantin, Tratat de statistică teoretică și economică, Editura Economică, București, 2008, ISBN 9789737093806

Fiecare nod al arborelui reprezintă un grup de obiecte ce se caracterizează prin trei indicatori: *(N, LS, SS)*, unde *N* este numărul de obiecte din cluster, *LS* şi *SS* au următoarele formule de calcul:

- pentru un nod frunză, *N*: $LS = \sum_{P_i \in N} P_i$ si $SS = \sum_{P_i \in N} |P_i|^2$;

- pentru un nod părinte (non-frunză), care are nodurile copil $N_1, N_2, \ldots N_k$: $LS = \sum_{i=1}^{k} LS$ al nodului N_i si $SS = \sum_{i=1}^{k} SS$ al nodului N_i.

Fiecare nod non-frunză conţine un anumit număr de copii. Numărul de copii pe care un nod non-frunză poate să îi aibă este limitat de un prag numit factor de adăugare. Diametrul unui subcluster dintr-un nod nu poate depăşi un anumit prag.

Inserarea unui obiect în arbore se realizează prin traversarea arborelui de sus în jos începând cu rădăcina în funcţie de distanţele euclidiană şi Manhattan. Obiectul este introdus în cel mai apropiat cluster sub un nod frunză. În cazul în care introducerea unui obiect într-un subcluster determină depăşirea diametrului şi implicit a pragului stabilit, atunci se creează un nou subcluster. Împărţirea unui nod frunză este efectuată de prima identificare a perechii de subgrupuri de sub nodul frunză care sunt separate de cea mai mare distanţa inter-cluster. Apoi toate celelalte subclustere sunt împărţite între cele două grupuri, având ca suport apropierea faţă de cele două subgrupuri.

1.3 Definire şi forma generală a unei reţele neuronale

La fel ca un creier uman o reţea neuronală este alcătuită din neuroni şi din conexiunile dintre ei (figura 3). În reţelele neuronale aceste conexiuni poartă denumirea de ponderi (weights). Informaţia electrică este simulată prin valori specifice stocate în aceste ponderi. Prin simpla schimbare a acestora se poate simula schimbarea structurii de conectare a reţelei.[13]

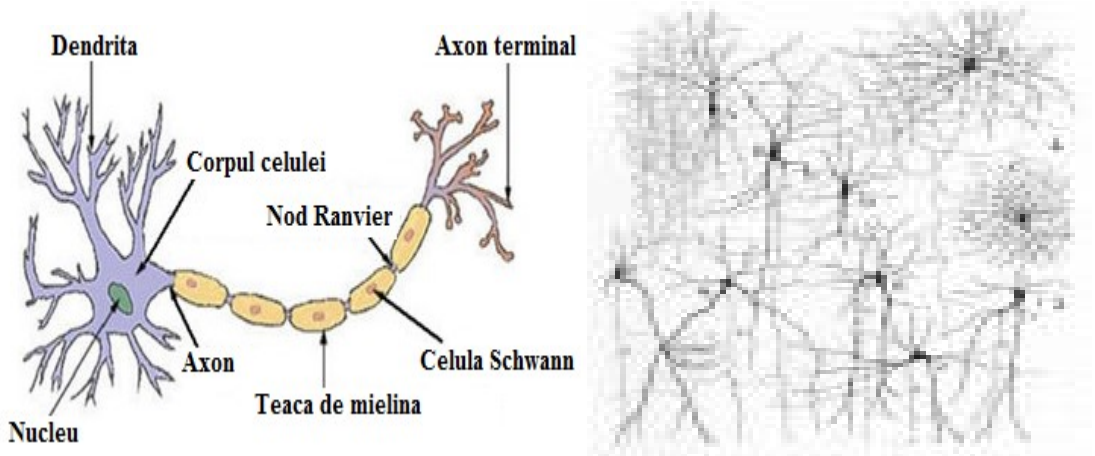

Figura 3. Reprezentarea neuronilor în creierul uman

[13] Caleanu Catalin- Daniel, Tiponut Virgil, Reţele Neuronale. Arhitecturi şi algoritmi, Editura Politehnica Timişoara, 2002, ISBN 973-9389-66-X