Categoria de servicii: ***WP 3. Servicii de dezvoltare a competențelor și de formare profesională (Skills Development and Training Services)***

Subcategoria de servicii: ***Sesiuni de instruire pe soluții digitale avansate (Training sessions on high performance digital solutions)***

SESIUNE DE INSTRUIRE PENTRU DATA MINING

Data mining – extragerea de cunoștințe din volume mari de date

Partener WEH: Universitatea *Spiru Haret*

Trainer de coordonare a transformarii digitale:

Andronie Mihai

CUPRINS

INTRODUCERE

Materialul de faţă tratează problema utilizării tehnicilor *data mining* în cadrul sistemelor informatice economice în vederea optimizării proceselor desfăşurate de organizaţiile economice.

Prezentul material este dedicat tehnicilor avansate de extragere a datelor din volume mari de date, cunoscute şi ca tehnici *data mining*. Se porneşte mai întâi de la apariţia şi dezvoltarea acestor tehnici odată cu dezvoltarea altor tehnologii cum ar fi aceea a bazelor de date. Sunt prezentate în continuare principalele tipuri de date care pot fi exploatate utilizând asemenea tehnici, printre acestea numărându-se datele stocate în baze de date, datele stocate în depozite de date, fluxurile de date, datele sub formă de serii de timp şi secvenţe, datele cu structuri de grafuri şi reţele şi, nu în ultimul rând, datele spaţiale, multimedia, text şi web. De asemenea sunt acoperite şi principalele funcţii *data mining* cum ar fi, de exemplu, asocierea, clasificarea, analiza excepţiilor, gruparea, regresia sau previziunile, cu diferite aplicaţii ale acestora în domeniul economic. La final este realizat un studiu comparativ referitor la câteva instrumente *data mining* disponibile pe piaţă astfel încât organizaţiile economice să aibă o imagine asupra acestora şi a diversităţii ofertei disponibile. Astfel, sunt acoperite diferite instrumente pentru *data mining* comerciale, produse de firme cum ar fi SPSS, Oracle, IBM sau Microsoft, fiecare având propriile caracteristici.

1. SCURTĂ PREZENTARE PRIVIND APARIŢIA TEHNICILOR DATA MINING ÎN PARALEL CU DEZVOLTAREA ALTOR TEHNOLOGII INFORMATICE

În urma importantelor progrese înregistrate în domeniul tehnologiei informaţiei, s-a ajuns la creşterea volumului de date produse şi stocate din diverse câmpuri de activitate ale societăţii umane, aceste date conţinând informaţii valoroase. De aici au rezultat în mod firesc diferite tehnici avansate de prelucrare a acestora, denumite sugestiv *data mining*.

Tehnologiile *data mining* răspund la nevoia de a prelucra aceste volume uriaşe de date pentru a obţine informaţii şi cunoştinţe utile în diverse domenii de activitate cum ar fi, de exemplu, analizele de piaţă, producţia, marketingul sau cercetarea ştiinţifică (Fig. 1).

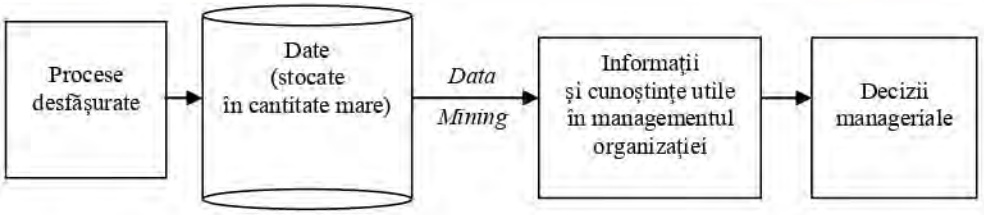

Fig. 1. *Rolul tehnicilor data mining în fundamentarea deciziilor manageriale*

Nu se poate concepe o analiză avansată a datelor fără o dezvoltare prealabilă a unor metode eficiente de stocare, de regăsire şi de prelucrare a acestora, lucru facilitat de tehnologiile din domeniul bazelor de date. Printre factorii care au contribuit la evoluţia bazelor de date relaţionale putem menţiona procesarea *on-line* a tranzacţiilor de date (OLTP – *On Line Transaction Processing*).

Astfel, îndată ce tehnologia a permis, în anii 1960-1970, au început să fie acumulate cantităţi din ce în ce mai mari de date, organizate pentru început în fişiere, apărând astfel primele baze de date.

Începând cu sfârşitul anilor '60, au apărut primele sisteme de gestiune a bazelor de date (SGBD sau DBMS – *Database Management Systems*). Acestea determinau îmbunătăţiri majore în domeniul bazelor de date, aducând în plus facilităţi cum ar fi limbajul de interogare SQL, posibilitatea de indexare a datelor pentru o accesare mai rapidă şi o optimizare a interogărilor, precum şi posibilitatea creării de rapoarte.

În anii '80 au apărut depozitele de date (*datawarehouse*) şi primele tehnici *data mining* destinate extragerii de informaţii şi cunoştinţe din date, continuând să se dezvolte până în prezent. Primele tehnici *data mining* au provenit din statistică, datorită faptului că, până să apară acestea, analiza datelor era făcută de statisticieni, care au început apoi să lucreze pe calculator folosind tehnici asemănătoare celor cunoscute de aceştia. Printre tehnicile *data mining* apărute se numără: clasificarea, asocierea, gruparea, analiza tendinţelor etc.

În acelaşi timp, s-a dezvoltat rapid internetul, ducând la apariţia de noi tipuri de date cum ar fi datele web, fluxuri de date etc. Concomitent cu acest fenomen, au fost înregistrate şi progrese semnificative în domeniul tehnologiei informaţiei care au permis stocarea datelor pe medii din ce în ce mai performante, precum şi o putere de calcul tot mai mare, care a deschis calea către o analiză mai avansată în vederea extragerii de cunoştinţe utile. Astfel, tehnicile *data mining* clasice au fost şi ele extinse pentru a putea exploata noile tipuri de date, acestea din urmă luând numele de *tehnici de nouă generaţie* şi fiind integrate de obicei în sistemele de gestiune a bazelor de date.

Au apărut apoi depozite de date (*data werehouse*) care pot conţine date eterogene aduse la un numitor comun în vederea exploatării lor ulterioare folosind diferite metode cum ar fi cele de

interogare, vizualizare sau chiar explorarea lor cu tehnici de tip OLAP (*On– Line Analytical Processing*), care permit analiza multidimensională a datelor. S-a ajuns astfel într-o situaţie în care se dispunea de date în cantităţi foarte mari, dar care nu puteau să fie valorificate. Din această cauză, au fost dezvoltate tehnici de analiză mai avansată a acestor date cum ar fi clasificarea, gruparea sau analiza excepţiilor, numite tehnici *data mining*, denumire sugestivă care asociază un proces familiar pentru toată lumea, mineritul (extragerea minereului din roci), cu extragerea cunoştinţelor din date.

Tehnicile *data mining* folosesc intensiv tehnologia informaţiei (IT) pentru analiza volumelor uriaşe de date care nu ar putea fi valorificate altfel tocmai datorită volumului mare de calcul necesar pentru această operaţie. În urma aplicării acestor tehnici, se obţin modele de date care pot fi valorificate în diverse domenii de activitate, printre care unul dintre cele mai importante este managementul strategic, unde este nevoie ca deciziile să fie fundamentate cât mai bine.

Noţiunea de *data mining* se referă în principiu la operaţia de extragere de cunoştinţe şi informaţii din volume mari de date [26], diferenţa principală între informaţii şi cunoştinţe fiind aceea că în timp ce primele au un pronunţat caracter de noutate, cunoştinţele se referă mai mult la ceva deja ştiut, relevant pentru un anumit domeniu. Ambele pot deveni date în momentul în care sunt înregistrate pe un mediu de stocare în vederea folosirii ulterioare. Este destul de greu să se găsească o definiţie exactă pentru acest câmp de activitate, deoarece cercetătorii tind să aibă opinii diferite în stabilirea unor limite exacte în care să fie încadrat acest concept. Astfel, există mai multe posibilităţi de definire, fiecare punând în lumină aspecte uşor diferite [37].

De-a lungul timpului au fost date mai multe definiţii pentru conceptul de *data mining*. Astfel, în 1992, Frawley, Piatetsky şi Matheus considerau că *data mining* reprezintă procesul de „extragere de informaţii netriviale, necunoscute anterior şi potenţial utile din date" [22]. Mai târziu, în 1999, procesul de *data mining* era considerat „procesul prin care informaţiile şi cunoştinţele sunt extrase din volume mari de date folosind tehnici care sunt mai mult decât o simplă căutare în date" [35].

Conform autorilor David Hand, Heikki Mannila şi Padhraic Smith [27], „*Data mining* este analiza unor seturi de date (de dimensiune mare), observaţionale, care are scopul de a identifica relaţii ascunse şi de a rezuma datele în moduri noi care sunt inteligibile şi utilizabile pentru proprietarul datelor". Alte lucrări [39], decât să dea o definiţie riguroasă noţiunii de *data mining*, preferă să explice contextul în care a apărut şi funcţia îndeplinită în diverse situaţii. Astfel, se poate spune că prin *data mining* se înţelege „aplicarea unor algoritmi cum ar fi arborii decizionali, gruparea, asocierea, analiza seriilor de timp, şi aşa mai departe, pe un set de date pentru a-i analiza conţinutul. Rezultatul analizei este reprezentat de modele, care pot fi explorate pentru a găsi informaţii valoroase" [39].

Termenul *data mining* este echivalent cu termenul descoperirea cunoştinţelor din date (*knowledge discovery from data – KDD*), acesta din urmă fiind mai exact, dar mai puţin sugestiv. De asemenea, acelaşi proces se mai poate „întâlni" cu denumiri precum extragerea cunoştinţelor (*knowledge extraction*), analiza datelor şi a modelelor sau chiar arheologia datelor (*data archeology*).

În urma celor enunţate anterior referitor la ce sunt tehnicile **data mining**, se poate considera că este important să se facă o distincţie clară între tehnicile de interogare aplicate asupra datelor disponibile (cum ar fi, de exemplu, interogările de tip SQL) şi tehnicile **data mining**. Din punctul de vedere al specialiştilor în domeniu, această distincţie este dată de faptul că în urma aplicării tehnicilor de interogare se obţin date care pot să apară în mod explicit în datele disponibile, în timp ce prin aplicarea de tehnici **data mining** se obţin date noi care nu apar ca atare niciunde în datele analizate. Ţinând cont de această observaţie. în continuare se va putea face o distincţie clară între interogări şi **data mining**.

În urma aplicării unor diferite tehnici şi procedee *data mining,* se obţin ca rezultate cunoştinţe care pot fi valorificate prin utilizarea lor la fundamentarea unor decizii manageriale, de marketing, sau pentru a stabili strategia pe termen lung a unei organizaţii economice.

Începând cu momentul apariţiei necesităţii de a extrage cunoştinţe din volume mari de date, aplicarea primelor tehnici *data mining* presupunea un număr de etape care trebuiau aplicate într-o ordine bine definită, care conduceau de la datele stocate la cunoştinţele extrase din ele. Acestea sunt [26], (Fig. 2.):

- curăţirea datelor prin înlăturarea acelora care nu sunt relevante sau care sunt inconsistente şi păstrarea acelora care conţin informaţie ce poate fi valorificată;
- integrarea datelor provenite din diferite surse eterogene într-un singur depozit de date omogen şi unitar;
- selectarea din depozitele de date disponibile a datelor relevante pentru a fi analizate ulterior;
- transformarea datelor în formate convenabile pentru analiză prin agregare, unificare etc.;
- procesul propriu-zis *data mining* prin care se extrag diferite modele ascunse din date;
- evaluarea modelelor obţinute şi păstrarea celor care sunt interesante şi utile;
- prezentarea cunoştinţelor obţinute către utilizator (manager în cazul unei organizaţii economice), folosind diferite modalităţi de reprezentare.

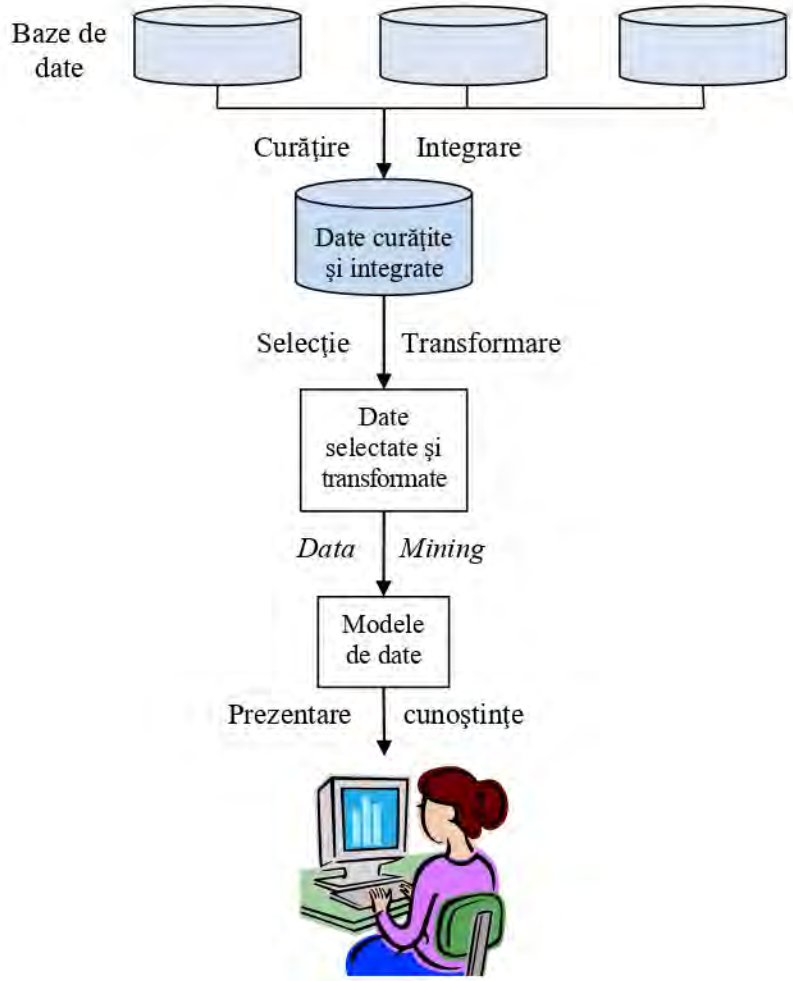

Fig. 2. *Etapele parcurse în cadrul procesului de extragere a cunoștințelor din date* [6]

2. EVOLUȚIA TEHNICILOR AVANSATE DE EXTRAGERE A DATELOR

Cu toate că domeniul *data mining* este unul relativ recent, în urma cercetărilor din ultimii ani au apărut o varietate foarte mare de algoritmi, metode și tehnici de lucru.

Un algoritm *data mining* este o procedură bine definită care preia date ca intrare și produce rezultate sub formă de modele [27]. Principala diferență dintre un algoritm și o metodă de calcul este aceea că algoritmul se termină după un număr finit de pași în timp ce metoda nu garantează că acest lucru se va întâmpla. Spre deosebire de acestea, conform *Dicționarului explicativ,* o tehnică este „un ansamblu de procedee și deprinderi folosite într-un anumit domeniu de activitate" [18].

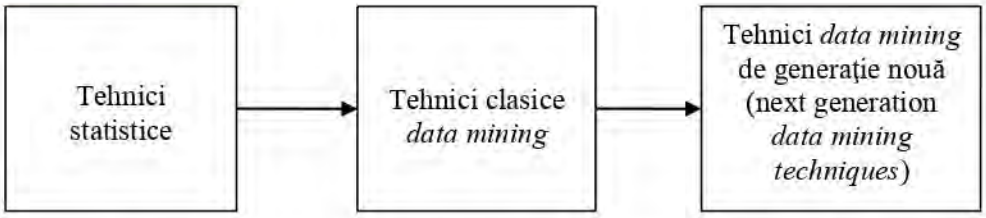

Fig. 3. *Evoluția de-a lungul timpului a tehnicilor* ***data mining***

Încă de la începutul dezvoltării *data mining*, au existat diferite metode și algoritmi de extragere a cunoștințelor utile din date. Primele tehnici *data mining* au derivat din metodele statistice și mai poartă numele de tehnici clasice (ca, de exemplu, *analiza statistică*, sau *analiza cluster*). Aceste tehnici au fost la început asociate cu colecțiile de date și au fost ulterior adaptate pentru a analiza volume uriașe de date (Fig. 3). Acestea rămân în continuare tehnicile cele mai utilizate *data mining* din cauză că specialiștilor în domeniu le sunt foarte familiare, ele fiind de obicei aplicate pe baze de date.

Odată ce tehnologia a evoluat și au apărut tipuri de date tot mai diverse, au început să se dezvolte și tehnicile *data mining* de generație nouă (*next generation techniques*), care sunt capabile să proceseze un volum de date incomparabil mai mare (față de primele baze de date și depozite de date), într-un timp relativ scurt. De asemenea, aceste tehnici pot procesa cu ușurință tipuri de date nestructurate care se dovedesc o sursă de date valoroasă, dar prea puțin exploatată. Astfel, putem prevedea că în viitor vor putea fi exploatate date cum ar fi conținutul lucrărilor care până acum erau preponderent în format tipărit – cărți (scanate în prealabil pentru a fi puse în format electronic), hărți etc. Totodată, ne putem aștepta ca internetul să fie exploatat la niveluri nemaiîntâlnite până acum, singurele piedici majore ce vor trebui depășite fiind gradul mare de eterogenitate a datelor de tip web și alegerea informațiilor utile și relevante dintre milioanele de pagini disponibile.

3. SURSE DE DATE CARE POT FI EXPLOATATE

3.1. Date stocate în baze de date

Bazele de date reprezintă una dintre formele cele mai uzuale de stocare a datelor și aplicarea asupra acestora a unor tehnici *data mining* corespunzătoare ar putea avea ca efect extragerea unor cunoștințe valoroase pentru domeniul de activitate unde acestea sunt folosite.

Bazele de date și depozitele de date au constituit suportul principal pentru aplicarea de tehnici avansate de analiză a datelor încă de la început, pe când celelalte surse de date, cum ar fi fluxurile de date, grafurile, datele multimedia, text și chiar web, au început să fie analizate ceva mai recent.

Bazele de date, dintre care cele mai utilizate sunt cele relaţionale, au reprezentat încă de la început o importantă sursă de date din care se puteau extrage cunoştinţe. Acestea reprezintă colecţii structurate de date, organizate în tabele, fiecare tabelă având un număr de coloane (atribute) şi un număr de linii (tupluri).

Bazele de date sunt create şi gestionate cu ajutorul unor sisteme de gestiune a bazelor de date (SGBD), utilizatorii neavând acces direct la fişierele care conţin datele. Aceste sisteme sunt proiectate astfel, încât să permită un acces facil, satisfăcând simultan mai mulţi utilizatori într-un timp oportun [17].

Pe bazele de date relaţionale există limbaje de interogare care permit regăsirea datelor cum ar fi SQL. Pe lângă bazele de date relaţionale, se pot aplica tehnici *data mining* şi pe alte tipuri de baze de date cum ar fi cele obiectuale, obiectual-relaţionale etc.

Este important să nu se confunde aplicarea de tehnici specializate *data mining* asupra datelor dintr-o bază de date cu interogările. În timp ce interogările doar regăsesc date din baza de date, tehnicile *data mining* pot să găsească modele ascunse, să identifice tendinţe, să facă predicţii sau să ofere informaţii care nu se regăsesc explicit undeva în baza de date analizată. Astfel, tehnicile *data mining* reprezintă o analiză mai avansată a datelor [15].

Bazele de date au avantajul că sunt structurate clar, ceea ce le face uşor de analizat, conţin informaţii exacte şi precise, fără ambiguităţi, şi sunt printre cele mai întâlnite surse de date disponibile, practic orice organizaţie modernă dispunând de o astfel de colecţie de date. Din acest motiv, bazele de date rămân în continuare unele dintre sursele cele mai valoroase de informaţii.

3.2. Date stocate în depozite de date

Depozitele de date sunt colecţii omogene de date obţinute în urma integrării mai multor surse de date, uneori diferite. Rezultatul nu trebuie neapărat să conţină toate informaţiile din bazele de date surse, ci poate să sumarizeze şi să reţină doar acele date care sunt considerate importante. Structura fizică a unui depozit de date poate fi relaţională sau sub formă de cub multidimensional [26].

Structura de cub multidimensional a unui depozit de date face posibilă şi aplicarea de tehnici de tip OLAP (*on-line analytical processing*). Acestea permit utilizatorului să vizualizeze datele la un grad de detaliere dorit. Prin operaţii OLAP datele pot fi generalizate (*operaţia de roll-up*), detaliate (*operaţia de drill-down*) sau, dacă se doreşte, se poate renunţa la anumite dimensiuni pentru a vizualiza mai bine ceea ce interesează.

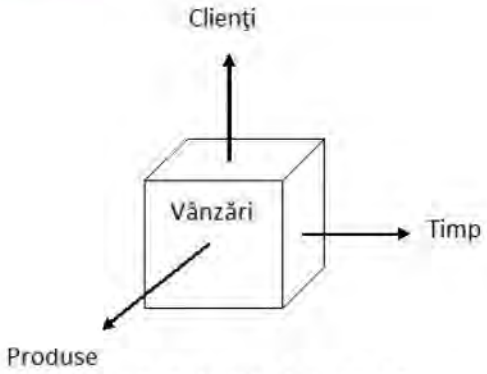

Fig. 4. *Cubul de date OLAP*

De multe ori, informațiile dintr-un depozit de date pot să fie vizualizate sub forma unui cub virtual cu ajutorul tehnicilor OLAP, datele putând să fie explorate în mai multe dimensiuni (Fig. 4).

Operațiile care pot fi realizate pe un cub de date OLAP sunt următoarele [36]:

- *Slice* – reprezintă un subset al cubului de date OLAP, astfel încât una din dimensiunile cubului (care să nu fie în acel subset) să ia o anumită valoare. Pentru cubul din Figura 4 se poate vedea, de exemplu, ce a cumpărat un client de-a lungul timpului sau ce clienți au cumpărat un anumit produs la momente diferite de timp;
- *Dice* – este o operație asemănătoare cu *Slice*, cu diferența că în acest caz se face pentru mai mult de două dimensiuni;
- *Drill down* – utilizatorul are posibilitatea de a mări nivelul de detaliu;
- *Drill up* – utilizatorul are posibilitatea de a reduce nivelul de detaliu, sumarizând;
- *Roll up* – operația de *roll up* presupune un calcul al relațiilor existente între datele disponibile pentru una sau mai multe dimensiuni;
- *Pivot* – operația de pivotare reprezintă o rotire a dimensiunilor datelor existente, astfel încât să se ajungă la o nouă prezentare a acestora; de exemplu, pentru cubul prezentat anterior, o pivotare poate fi trecerea de la produse în funcție de clienți la clienți în funcție de produse.

3.3. Fluxuri de date

Aplicarea tehnicilor avansate *data mining* pe fluxurile de date are un potențial uriaș pentru situații cum ar fi analiza traficului pe internet, a datelor generate de sistemele de supraveghere în timp real, a datelor generate în urma proceselor de producție industrială și a altor tipuri de date generate de activități cu o dinamică însemnată [26].

Conform aceleiași surse, fluxurile de date sunt cantități potențial infinite de date primite sau transmise de către un sistem informatic și care necesită o serie de tehnici specializate pentru a putea

fi analizate în timp real. Din această cauză, cât și din cauza faptului că fluxurile de date sunt de obicei prea mari pentru a fi stocate undeva, algoritmii și metodele specifice au de prelucrat această cantitate de date folosind o singură parcurgere. De asemenea, datele conținute într-un astfel de flux se pot schimba rapid și algoritmii nu au la dispoziție un timp nelimitat pentru a găsi soluția optimă.

Este un lucru obișnuit ca tehnicile *data mining* pentru fluxuri de date să analizeze numai un eșantion din tot fluxul pentru a găsi o soluție mai rapid chiar dacă aceasta nu este cea mai exactă (Fig. 5).

De exemplu, dacă o instituție economică are de luat o decizie și aceasta se dorește să fie fundamentată pe analiza datelor dintr-un flux, de cele mai multe ori nu este practic să se aștepte găsirea soluției optime de către sistemul de calcul. Poate fi o abordare mai profitabilă aceea de a găsi mai rapid o soluție chiar dacă aceasta nu este cea optimă.

Pentru lucrul cu fluxuri mari de date au fost create sisteme de gestiune a fluxurilor de date (*Data Streams Management Systems – DSMS*), care pot gestiona la un moment dat chiar mai multe fluxuri. O particularitate a datelor din fluxurile de date este aceea că o secvență de date se va pierde dacă nu este memorată undeva pentru a fi analizată ulterior, iar un sistem de gestiune a fluxurilor de date trebuie să țină cont de acest aspect. Cu un astfel de sistem, se pot realiza interogări pe fluxuri, care sunt de două feluri:

- interogări realizate o singură dată; acestea sunt formulate de către utilizator și evaluate o singură dată, oferind răspunsuri bazate pe starea fluxului de date la un singur moment de timp;
- interogări continue; sunt acele interogări evaluate în permanență după ce acestea au fost formulate de către utilizator și rezultatele lor variază în timp odată cu variația datelor din fluxul analizat.

Pentru ca un sistem de gestiune a fluxurilor de date să poată să ofere rezultate exacte la interogări pe baza unui flux de date, acesta trebuie să ruleze pe un sistem de calcul cu o cantitate practic nelimitată de memorie și cu o putere de procesare foarte mare, lucru care de obicei nu se întâmplă. Din această cauză, în mare parte din cazuri, rezultatele obținute pe baza unor interogări sunt oarecum aproximative, precizia acestora fiind limitată de performanța sistemului de calcul utilizat.

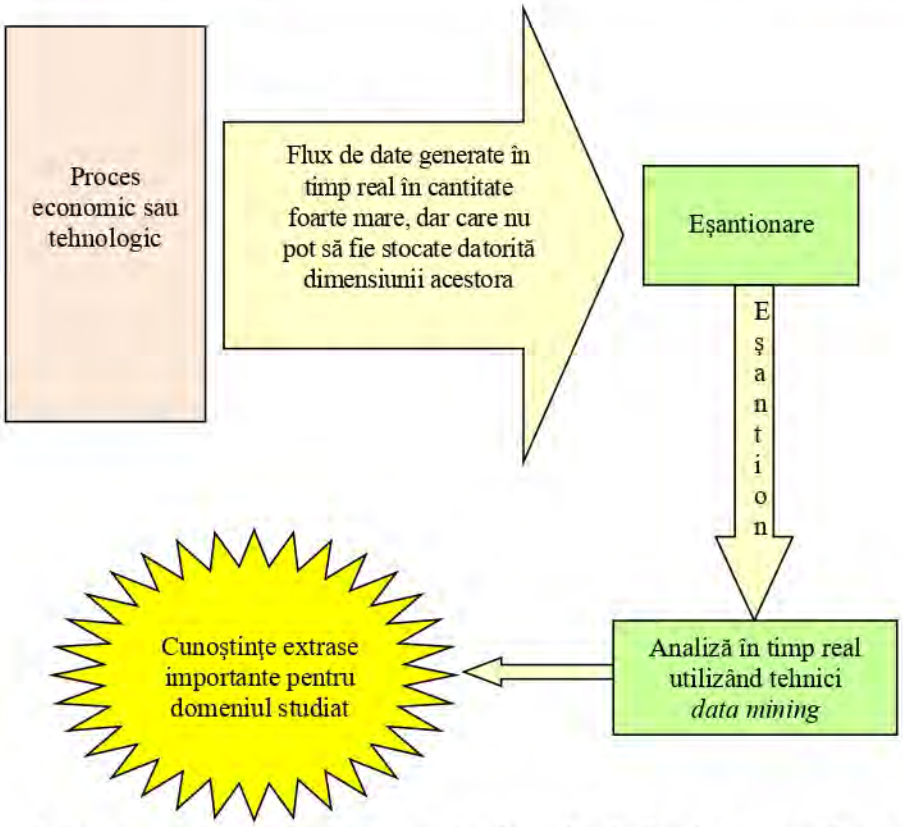

Fig. 5. *Eşantionarea şi analiza în timp real a fluxurilor de date în cantitate foarte mare*

Exemple de tehnici şi metode aplicate frecvent pe fluxurile de date sunt histogramele, ferestre mobile (*sliding windows*), metodele de reducere a datelor, extragerea de eşantioane aleatoare etc. [30]. Tot în lucrarea menţionată se afirmă că toate aceste tehnici contribuie la analiza fluxurilor de date complexe multidimensionale şi sunt integrate în sisteme de management al fluxurilor de date.

O altă abordare posibilă în cazul fluxurilor de date este folosirea unor tehnici similare cu OLAP adaptate pentru fluxuri de date (*Stream OLAP*), care încearcă să prezinte datele componente ale unui flux ca pe datele dintr-o bază de date relaţională, astfel încât utilizatorul final să poată alege nivelul de abstracţie care îl doreşte. Aplicarea de tehnici *Stream OLAP* este avantajoasă, pentru că, în multe cazuri datele dintr-un flux sunt multidimensionale şi, în consecinţă, este nevoie de o analiză multidimensională a acestora. Utilizatorul ar trebui să poată în acest caz să efectueze operaţii de *drill-down/up* pentru a putea vizualiza datele atât în detaliu, cât şi mai puţin detaliat (mai sintetic) [36].

3.4. Serii de timp şi secvenţe

O bază de date cu serii de timp constă în secvenţe de „valori sau evenimente obţinute în urma unor măsurători repetate de-a lungul timpului" [30]. Tot aceeaşi sursă specifică faptul că diferenţa între datele de tip serie de timp şi datele secvenţiale este că, deşi ambele lucrează cu elemente ordonate, în cazul particular al datelor de tip serie de timp este vorba de o ordonare temporală.

Exemplele de date tip serie de timp întâlnim frecvent, cum ar fi tranzacţiile de pe piaţa bursieră, rezultatele obţinute în urma unor experimente ştiinţifice, variaţia unor indicatori de natură economică etc.

Datele din bazele de date cu serii de timp pot fi măsurate la intervale regulate, dar acest lucru nu este obligatoriu, acestea putând fi obţinute şi la perioade neregulate de timp cu condiţia să fie în ordine cronologică.

Ca şi fluxurile de date, şi datele serii de timp pot să fie prezente în cantităţi mari şi, prin urmare, este nevoie de algoritmi performanţi pentru a le putea analiza.

Un exemplu de analiză a datelor de tip serie de timp se găseşte în partea aplicativă a acestei lucrări, în ultimul capitol. Acolo este prezentată şi sub formă grafică o astfel de analiză. Pe un astfel de grafic se pot evidenţia tendinţele, iar prin utilizarea de tehnici avansate de analiză a datelor se pot face chiar şi previziuni pe termen mai scurt.

3.5. Date cu structură de grafuri şi reţele

Extragerea cunoştinţelor din grafuri şi reţele sociale este o altă direcţie de dezvoltare în domeniul *data mining*. Grafurile sunt importante pentru că permit modelarea unor structuri eterogene, cum ar fi datele de tip text, datele de tip web, reţele sociale, şi a altor date care prezintă un grad ridicat de complexitate. Extragerea de cunoştinţe din baze de date conţinând grafuri „poate fi văzută ca o căutare în mulţimea tuturor subgrafurilor posibile" [21]. Astfel de tehnici sunt variate, dintre ele putând să enumerăm *abordarea apriori*, abordarea creşterii progresive a modelului (*Pattern-Growth approach*) etc.

Grafurile îşi au originea în matematică, structurile de date sub formă de grafuri găsindu-şi originea acolo. Un graf constă într-un număr finit de noduri şi un număr finit de muchii care unesc nodurile. Grafurile pot fi, de asemenea, orientate sau neorientate.

Pentru a putea reprezenta un graf în memorie, de obicei se utilizează structuri de date sub formă de matrici. Dacă graful este neorientat, matricea sa de incidenţă va fi simetrică; în caz contrar, aceasta nu va fi simetrică.

Utilitatea grafurilor este aceea că sub formă de grafuri pot să fie reprezentate date referitoare la multe domenii de activitate umană.

Spre exemplu, nodurile unui graf pot fi orașe, iar muchiile drumurile între acestea. Muchiile pot avea ponderi diferite, ceea ce ar putea corespunde cu distanțe diferite între orașe. De asemenea, dacă graful ar fi orientat, ar însemna că pe anumite drumuri se circulă doar într-un anumit sens.

Rețelele sociale sunt „seturi de date eterogene și multirelaționale reprezentate sub formă de grafuri", care sunt folosite pentru a modela situații întâlnite în lumea reală cum ar fi convorbirile telefonice într-o rețea, răspândirea virușilor sau comunicațiile prin internet [1].

Analiza datelor cu structură de rețele sociale se face similar cu analiza datelor cu structură de grafuri [42].

3.6. Date spațiale, multimedia, text și web

Noile evoluții din domeniul tehnologiei informației au făcut posibilă extragerea de cunoștințe din noi tipuri de date, mai puțin convenționale.

Există baze de date speciale, care sunt destinate stocării diferitor date spațiale cum ar fi, de exemplu, imagini, informații cartografice, hărți etc. [26]. În mod similar ca și pentru bazele de date relaționale, este posibil și în acest caz să se construiască depozite de date spațiale (*data warehouse*) care să fie explorate cu tehnici similare cu cele OLAP (*Spatial OLAP*). De exemplu, putem avea într-o bază de date hărți economice, cum ar fi o hartă cu un anumit indicator economic (exemplu: venitul pe locuitor) peste care să aplicăm tehnici de detaliere, mergând cu informația până la nivel de comună, sau generalizând până la nivel de județ sau la nivel național, acest lucru fiind posibil prin folosirea unor tehnici similare cu OLAP.

În mod asemănător cu bazele de date spațiale, există baze de date care conțin date multimedia, stocând un volum imens de date. Și în acest caz, este posibil să fie construit un cub de date multidimensional similar cu cel din cazul bazelor de date relaționale și să fie efectuată o analiză multidimensională a datelor multimedia [26]. Pe lângă acest tip de analiză, mai pot fi extrase și alte cunoștințe din datele multimedia, cum ar fi, de exemplu, analiza imaginilor sau filmelor pentru recunoașterea feței, în cazul aplicațiilor pentru securitate.

Un alt tip de date care pot fi analizate sunt datele de tip text, care, în format electronic, au devenit tot mai numeroase în ultimii ani, tehnicile *data mining* pe text dezvoltându-se corespunzător. Cu toate că aplicarea tehnicilor *data mining* pe documente, cărți și alte texte în format electronic are rolul de a extrage cunoștințe, nu trebuie să confundăm acest proces cu simpla regăsire a informației în acestea [29]. Aplicarea tehnicilor *data mining* pe text poate avea și aspecte mult mai complexe. În mod obișnuit, datele de tip text sunt nestructurate și necesită tehnici speciale pentru a putea fi analizate.

Odată cu dezvoltarea foarte rapidă a internetului, acesta a devenit una dintre cele mai importante surse de date disponibile [12]. În consecință, au început să apară diferite tehnici *data*

mining pentru extragerea cunoștințelor de pe web care au un potențial mare de folosire. Conform autorilor Han J. și Kamber M., aplicarea acestor tehnici pe web se confruntă cu anumite probleme specifice cum ar fi:

- imensa cantitate de informație disponibilă;
- complexitatea ridicată a paginilor web;
- dinamica foarte ridicată a informațiilor de pe internet;
- faptul că nu toate aceste informații sunt relevante.

Datele folosite pentru experimentare sunt stocate într-o bază de date relațională, descrisă în detaliu la finalul lucrării de față.

4. FUNCȚII DATA MINING

Tehnicile, algoritmii și metodele *data mining* au fiecare un scop pentru care au fost proiectate, fiecare realizând o anumită funcție *data mining*. Printre cele mai frecvente funcții pe care le întâlnim la sistemele *data mining* sunt gruparea (*clustering*), asocierea, clasificarea, regresia, previziunile, analiza secvențelor de date (serii de valori discrete), analiza excepțiilor (*outlyer analysis*) (Fig. 6).

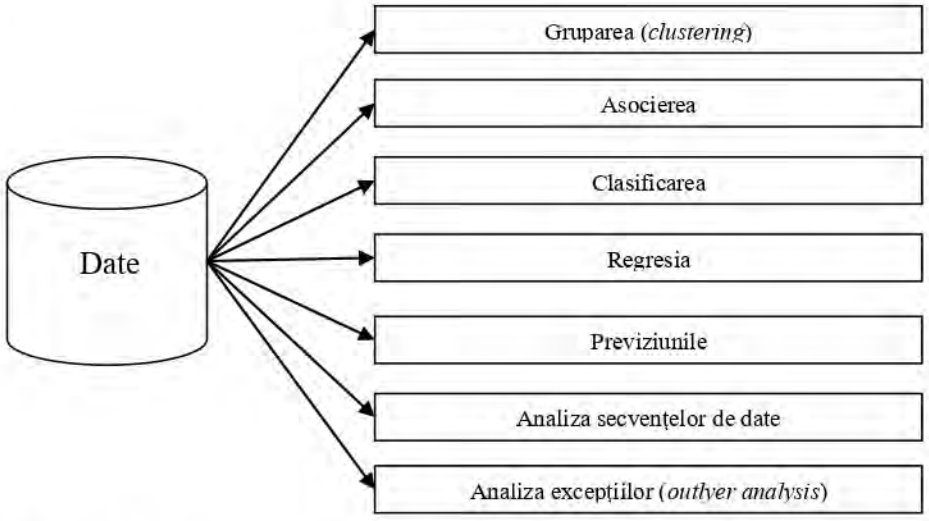

Fig. 6. *Principalele funcții* ***data mining*** *care se pot aplica pe o cantitate importantă de date*

În Tabelul 1 sunt prezentate pe scurt aceste funcţii *data mining,* subliniindu-se specificul şi caracteristicile fiecăreia. Multe dintre ele se aseamănă într-o anumită măsură, dar există şi diferenţe care le delimitează.

Astfel, de exemplu, se poate spune că atât clasificarea, cât şi gruparea au ca efect împărţirea „obiectelor" în anumite clase, dar diferenţa este modul în care această împărţire se realizează, la clasificare clasele fiind prestabilite, pe când la grupare, fiind construite ţinând cont de caracteristicile „obiectelor" analizate.

O altă diferenţă care merită amintită este aceea dintre asociere şi analiza secvenţelor. Dacă prima, pur şi simplu, constată nişte reguli şi modele în datele analizate, cea din urmă caută nişte modele pe nişte date ordonate în timp, încercând să determine şi o ordine de desfăşurare în timp a acestor modele, între evenimente presupunând că există anumite implicaţii.

De asemenea, trebuie făcută o diferenţă între regresie şi funcţia de previziune. Regresia trasează o tendinţă generală care, dacă se desfăşoară în timp, poate fi folosită, printre altele, şi la previziune, pe când în cazul funcţiei de previziune lucrurile pot fi mai complexe putând interveni şi alţi factori, putând ţine cont de ciclicitatea unor evenimente etc.

La rândul său, şi analiza excepţiilor poate fi înrudită cu alte funcţii cum ar fi gruparea sau clasificarea, acestea din urmă neavând scopul de a găsi excepţiile, dar putând fi utilizate şi pentru aşa ceva. Astfel, dacă se grupează „obiectele" pe categorii în funcţie de caracteristicile lor, se va putea constata uşor că anumite grupe sunt formate dintr-un singur „obiect" foarte diferit de celelalte, acesta putând fi considerat excepţie. La fel se poate întâmpla cu clasificarea când un „obiect" nu se încadrează în nicio clasă. Pe lângă acestea, există tehnici şi algoritmi specializaţi pentru identificarea şi analiza excepţiilor.

Tabelul 1

Principalele funcţii ***data mining*** [39]

Funcţia *data mining*	**Caracteristici**	**Exemple**
Gruparea (*clustering*)	• Ajută la găsirea unor categorii de „obiecte", bazându-se pe atributele acestora; • Grupurile formate cuprind „obiecte" asemănătoare din mai multe puncte de vedere [24]; • Algoritmii pentru grupare lucrează în mod nesupervizat, luând la intrare atributele obiectelor şi întorcând grupurile obţinute.	• Gruparea clienţilor unei firme în funcţie de valoarea achiziţiilor acestora; • Gruparea produselor după caracteristicile acestora.
Asocierea	• Are rolul de a identifica seturi frecvente de obiecte care apar împreună; • Uneori, utilizatorul poate alege ce obiecte să fie analizate sau cât de frecvente să fie acestea;	• Identificarea de produse care se vând cel mai frecvent împreună (Analiza coşului de cumpărături);

Funcția *data mining*	Caracteristici	Exemple
	• Poate indica și reguli de asociere cu probabilitatea ca acestea să se confirme.	• Identificarea zonelor în care se vând mai mult anumite produse; • Identificarea perioadelor de timp în care vânzările cresc.
Clasificarea	• Este una din cele mai frecvente activități *data mining*, răspunzând la probleme cum ar fi analiza riscului [31]; • Spre deosebire de grupare, unde categoriile nu sunt prestabilite, aici se introduc obiectele în categorii prestabilite, în funcție de caracteristicile acestora; • Algoritmii de clasificare lucrează în mod supervizat, necesitând niște criterii pentru a împărți obiectele [19]; • De obicei, criteriile de clasificare trebuiesc stabilite anterior, ca rezultat al analizei datelor istorice deținute; • Printre tehnicile de clasificare se pot include arborii decizionali, rețelele neurale etc.	• Clasificarea clienților unei bănci în funcție de cum și-au plătit ratele; • Analiza riscului în asigurări în vederea stabilirii primei de asigurare pentru diverse cazuri concrete; • Stabilirea impozitului pe proprietate în funcție de anumite criterii.
Regresia	• Provine din statistică, unde a fost folosită intens; • Este similară clasificării, cu diferența că lucrează pe atribute cu valori continue [34]; • Poate fi folosită pentru predicție.	• Diferite predicții liniare sau polinomiale (neperiodice); • Calculul unor valori intermediare prin interpolare.
Previziuni	• Are rolul de a face o previziune în timp, ținând cont de obicei de valorile din trecut sub formă de serii de evenimente desfășurate în timp; • Spre deosebire de regresie, poate lua în calcul și alți factori cum ar fi periodicitatea unor evenimente.	• Previziuni sezoniere în diferite domenii de activitate; • Previziuni ale vânzărilor în vederea aprovizionării cu marfă.
Analiza seriilor de timp și a secvențelor	• E folosită pentru identificarea de modele în serii de valori discrete; • Diferența dintre analiza secvențelor și asociere este că prima se ocupă și de tranzițiile și ordinea de trecere între diferite stări în timp ce a doua doar caută asocieri între obiecte presupuse independente [5].	• Identificarea de modele urmate de vânzările unei firme de-a lungul timpului.
Analiza excepțiilor (*outlyer analysis*)	• Are rolul de a detecta dintr-un număr de „obiecte", pe acelea care se comportă foarte diferit de restul.	• Este folosită pentru detecția fraudelor, analiza intruziei, detecția erorilor etc.

Primele tehnici și metode *data mining* apărute, denumite și tehnici *data mining* clasice, lucrau pe date provenite din colecții structurate, cum ar fi bazele de date sau depozitele de date. Pe lângă acestea, mai există și o altă categorie de tehnici, cunoscute ca tehnici de generație nouă (*next*

generation techniques). Acestea au apărut ceva mai târziu, ca rezultat al cercetărilor în domeniul bazelor de date şi *data mining* şi au scopul de a lucra pe seturi de date eterogene cum ar fi fluxurile de date sau web-ul. Trebuie menţionat faptul că funcţiile specificate în Tabelul 1 pot fi aplicate atât de tehnicile clasice *data mining,* cât şi de tehnicile de generaţie nouă. Astfel, primele tehnici folosesc în majoritatea cazurilor baze de date sau depozite de date care sunt destul de bine structurate, în timp ce tehnicile de nouă generaţie sunt gândite pentru a fi aplicate pe o varietate mult mai mare de date nestructurate şi eterogene.

5. ALGORITMI DATA MINING

De-a lungul timpului, au fost dezvoltaţi un număr foarte mare de algoritmi *data mining*, mai mult sau mai puţin complecşi, implementând de obicei una dintre funcţiile *data mining* descrise la capitolul 4 [40].

Conform *Dicţionarului explicativ al sistemelor de baze de date*, un algoritm este un „set finit de reguli care dă o secvenţă de operaţii pentru soluţionarea unui tip specific de probleme" [41]. Un algoritm *data mining* este văzut ca o succesiune finită de paşi (etape) care pot fi parcurşi pentru a extrage informaţii valoroase ce nu apar în mod explicit într-un volum mare de date şi nu ar putea să fie extrase prin interogări. Ca şi orice alt algoritm, un algoritm *data mining* trebuie să îndeplinească cel puţin următoarele trei criterii:

- să fie finit – orice algoritm trebuie să se termine la un moment de timp; din acest motiv, trebuie ca aceia care proiectează un algoritm să se asigure că, după un număr finit de paşi, se va ajunge la soluţia problemei, iar aceia care implementează un algoritm într-un limbaj de programare trebuie să se asigure că este implementat corect, astfel încât să nu existe posibilitatea de intrare în bucle infinite;
- să fie determinist – fiecare pas al unui algoritm trebuie să fie precis definit, să se cunoască la fiecare pas al acestuia exact ce trebuie făcut şi care va fi pasul următor;
- să primească date de intrare şi, pe baza acestora, să obţină date de ieşire, acestea reprezentând, de fapt, soluţia problemei; în cazul algoritmilor de *data mining*, datele de intrare reprezintă, de obicei, datele de analizat, iar datele de ieşire sunt reprezentate de informaţiile şi cunoştinţele obţinute la final pe baza acestora;

În ordine cronologică, primii algoritmi *data mining* apăruţi au fost cei clasici, evoluaţi din tehnicile clasice de analiză a datelor, bazate în mare parte pe statistică. Abia mai târziu au apărut algoritmii de nouă generaţie, mai performanţi, mai evoluaţi şi capabili de a analiza şi date în format neconvenţionali, folosindu-se arbori decizionali pentru a îndeplini funcţii *data mining* cum ar fi clasificarea, previziunea etc [44].

În continuare, vom prezenta câțiva dintre cei mai cunoscuți algoritmi *data mining* cu mențiunea că pe lângă aceștia mai există alte câteva mii de algoritmi de analiză a datelor mai puțin cunoscuți, unii dintre aceștia fiind proiectați pentru probleme specifice. Fiecare dintre algoritmii prezentați în continuare a fost proiectat pentru a îndeplini o anumită funcție *data mining*.

5.1. Algoritmul k-means

K-means este un algoritm euristic care realizează funcția *data mining* de grupare (*clustering*). Acest algoritm a apărut încă din anii '60 ai secolului trecut, fiind denumit *k-means* încă din 1967, de către James MacQueen [20].

Algoritmul *k-means* funcționează în mod iterativ și se aplică pe un număr finit de obiecte, grupându-le în categorii (*clustere*), formate ad-hoc, în funcție de caracteristicile acestora. În cadrul algoritmului *k-means* fiecare *cluster* este caracterizat pe baza centrului acestuia (văzut în sens geometric, ca o medie a caracteristicilor obiectelor din *clusterul* respectiv), ca în Figura 7.

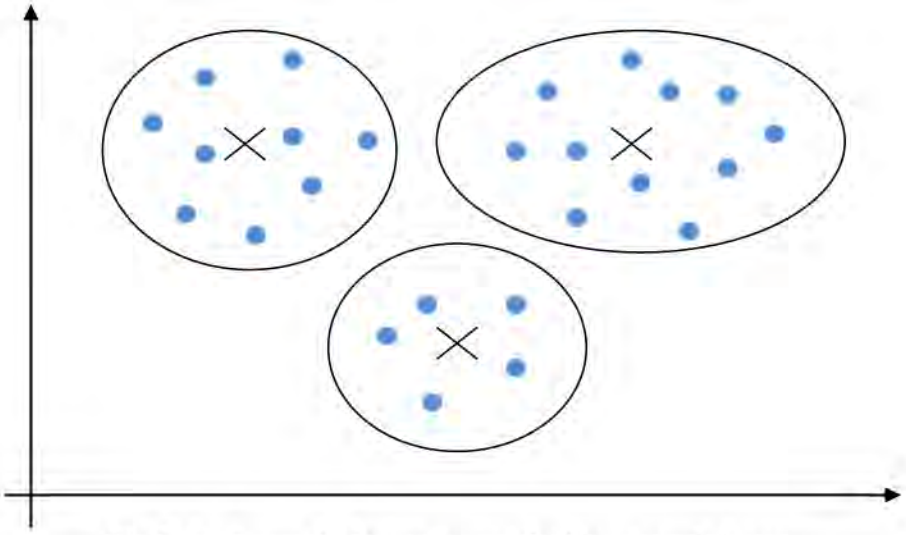

Fig. 7. *Realizarea de **clustere** (grupe) de obiecte asemănătoare într-un spațiu cu două dimensiuni*

În cadrul algoritmului *k-means*, fiecărui obiect îi este asociat un identificator care să arate cărui *cluster* acesta îi aparține. K reprezintă numărul total de *clustere* obținute, iar centrele acestora vor fi calculate ca o medie a caracteristicilor obiectelor care formează fiecare *cluster*.

Algoritmul *k-means* are ca intrări numărul de *clustere* care se dorește a fi identificate, notat cu k, și mulțimea de obiecte care trebuie împărțite în *clustere*, cu caracteristicile specifice ale acestora.

Ca ieşire a acestui algoritm vor fi coordonatele centrelor celor *k clustere* obţinute şi un set de identificatori care arată fiecare obiect în ce *cluster* va fi încadrat.

Algoritmul *k-means* are doi paşi, care se repetă:

Pasul 1: Alocarea obiectelor în cele mai apropiate *clustere* în funcţie de caracteristicile acestora (pe baza distanţei dintre centrul *cluster*-ului şi obiect);

Pasul 2: Recalcularea centrelor *clusterelor* în urma încadrării în acestea de noi obiecte (ca medie a caracteristicilor obiectelor componente).

Condiţia de oprire din iteraţie este convergenţa algoritmului (când se ajunge la o situaţie stabilă).

Algoritmul *k-means* are atât avantaje, cât şi limite. Este un algoritm care a fost îmbunătăţit de-a lungul timpului odată cu evoluţia tehnicilor *data mining* şi se bucură încă de succes.

5.2. Algoritmul C4.5

C4.5 este un algoritm *data mining* care realizează funcţia de clasificare. Algoritmul C4.5 este văzut şi ca un algoritm de învăţare, acesta luând ca parametri de intrare nişte exemple [33]. Pe baza acestora se calculează un arbore de decizie, folosit ulterior pentru clasificarea altor obiecte, fiecare după caracteristicile sale. În crearea unui arbore de decizie, la început acesta este construit după exemplele deţinute, urmând ca ulterior, pe baza unor exemple suplimentare să fie simplificat.

Algoritmul C 4.5 execută recursiv următorii paşi:

Pasul 1: Se iniţializează cu mulţimea vidă arborele de decizie care va fi returnat;

Pasul 2: Pentru fiecare caracteristică a unui obiect din exemplele de învăţare se calculează câştigul de informaţie dacă s-ar lua ca criteriu de ramificare a arborelui;

Pasul 3: Se ia acel atribut cu câştigul de informaţie maxim;

Pasul 4: Se creează un nod rădăcină în arborele de decizie unde ramificaţia se face după acest atribut;

Pasul 5: Se împarte setul iniţial în două subseturi în funcţie de atributul ales;

Pasul 6: Se execută recursiv procedura anterioară pentru cele două subseturi obţinute la pasul 5;

Pasul 7: Se adaugă cei doi subarbori obţinuţi în urma pasului 6 la arborele de decizie care va fi returnat.

Ca şi alţi algoritmi *data mining*, şi C 4.5 a fost îmbunătăţit de-a lungul timpului, următoarea versiune a acestuia fiind See5/ C 5.0. O versiune a algoritmului C 4.5 a fost descrisă şi de Freund şi Schapire, aceasta prezintă o eficienţă mai ridicată, precum şi o scalabilitate mai bună [23].

5.3. Algoritmul Apriori

Apriori este un algoritm *data mining* care realizează funcţia de asociere. Acest algoritm descoperă tipare comune în datele analizate într-un număr finit de paşi. Acest algoritm a fost dezvoltat în anul 1994 de cercetătorii Agrawal şi Srikant, bazându-se pe identificarea de seturi candidate în vederea identificării seturilor frecvente de date [38].

Denumirea algoritmului *Apriori* provine de la faptul că acesta utilizează cunoştinţe anterioare referitoare la proprietăţile seturilor frecvente [26], bazându-se pe proprietatea *Apriori*, care afirmă că toate subseturile nevide ale unui set frecvent sunt, de asemenea, frecvente la rândul lor. În algoritmul *Apriori* se ia un prag limită de jos denumit *minsup*, iar dacă un set nu are mai mult de un *minsup* apariţii acesta nu poate să fie considerat set frecvent şi, prin urmare, nici toate subseturile acestuia nu pot fi considerate a fi frecvente [13].

Algoritmul *Apriori* se bazează pe parcurgerea iterativă a următorilor doi paşi pentru calcularea unei mulţimi de subseturi de k elemente, bazându-se pe o mulţime de subseturi frecvent de k-1 elemente:

Pasul 1: constă în reunirea cu ea însăşi a mulţimii se subseturi de k-1 elemente, obţinând astfel o mulţime de seturi de elemente candidat; se obţin la acest pas o mulţime de seturi de k elemente, dintre care nu toate sunt frecvente (unele nu au numărul de apariţii mai mare de *minsup*);

Pasul 2: constă în curăţarea mulţimii obţinute astfel încât să rămână doar acele subseturi care apar de mai mult de *minsup* ori; în acest pas se utilizează proprietatea *apriori* pentru a reduce volumul de calcul.

Ulterior, au apărut diferite variante ale algoritmului *Apriori*, dintre acestea amintind *AprioriTID*, care oferă performanţe mai bune la valori mai mari ale lui k, deşi la valori mici totuşi *algoritmul apriori clasic* dă performanţe mai bune. O altă îmbunătăţire a algoritmului *Apriori* o reprezintă algoritmul *Frequent Pattern Growth,* care a fost descris de către J. Han, J. Pei şi Y. Yin în lucrarea [28].

5.4. Algoritmul k Nearest Neighbour

Algoritmul k Nearest Neighbour (pe scurt kNN) este un algoritm de clasificare care împarte un număr de obiecte în clasele în care se încadrează obiectele cu caracteristicile cele mai asemănătoare de acestea [16].

Principiul acestui algoritm este ca atunci când se dispune de un obiect nou şi se doreşte încadrarea acestuia într-una din clasele existente, se iau cei mai apropiaţi k vecini ai acestui obiect şi se clasifică obiectul în funcţie de apartenenţa acestor vecini. Pentru a determina cele k obiecte

cele mai apropiate, este necesar să se calculeze distanţa de la noul obiect la toate celelalte obiecte care sunt încadrate deja în clase, pe baza atributelor acestora (care sunt cunoscute) şi să fie reţinute cele cu distanţele cele mai mici. În funcţie de apartenenţa obiectelor asemănătoare la clasele existente, se va decide în ce clasă să fie încadrat noul obiect.

Ca intrare, *algoritmul kNN* ia o mulţime de clase, un set de obiecte de instruire care sunt încadrate în clase cunoscute, precum şi un obiect nou care va trebui atribuit unei clase în funcţie de caracteristicile acestuia. Ca ieşire, algoritmul returnează clasa în care este încadrat noul obiect.

Paşii pe care *algoritmul kNN* îi presupune sunt următorii:

Pasul 1: se calculează distanţa de la obiectul nou la obiectele din setul de instruire;

Pasul 2: se identifică obiectele cu distanţele cele mai mici faţă de noul obiect (vecinii cei mai apropiaţi);

Pasul 3: se asociază noul obiect unei clase, după cum sunt majoritatea vecinilor acestuia.

K Nearest Neighbour este un algoritm popular, care a fost implementat în multe limbaje de programare de-a lungul timpului [43].

În afară de algoritmii *data mining* descrişi anterior, există şi alţi algoritmi, unii dintre aceştia fiind destinaţi pentru diferite situaţii particulare. Unii dintre aceşti algoritmi pot prezenta informaţiile obţinute şi sub formă grafică, fiind în consecinţă mai uşor de înţeles de către nespecialişti. Astfel, în studiul de caz de la finalul lucrării noastre au fost implementaţi algoritmi care să poată fi aplicaţi pe cazul particular prezentat în lucrare şi rezultatele acestora pot fi vizualizate şi grafic.

6. PREZENTARE COMPARATIVĂ A CÂTORVA INSTRUMENTE DATA MINING

6.1. Criterii de evaluare a instrumentelor data mining

Cu toate că până în prezent au apărut un număr foarte mare de algoritmi care implementează tehnici *data mining*, numai o mică parte dintre aceştia au fost utilizaţi de către produsele software comerciale disponibile şi, prin urmare, pot fi aplicaţi pentru rezolvarea unor probleme de afaceri reale [32]. După cum afirmă Gupta, Bhatnagar şi Wasan, există bariere care împiedică ultimele tehnologii apărute în domeniul *data mining* să fie folosite în rezolvarea unor probleme din lumea reală, cum ar fi, de exemplu, necesitatea ca utilizatorul tehnicilor *data mining* să înţeleagă procesul care se derulează şi să interpreteze corect rezultatele obţinute [25].

Întreprinderile au la dispoziție o mare varietate de produse comerciale *data mining* și trebuie să le evalueze și să aleagă dintre acestea pe care cele se potrivesc cel mai bine pentru tipul de date de care dispun și care ajută cel mai mult la rezolvarea problemelor pe care le întâmpină. Pentru că diferitele produse *data mining* tind să fie foarte diferite între ele în majoritatea aspectelor, după cum precizează Jiawei Han și Micheline Kamber, o companie ar trebui să ia în considerare un număr de criterii în funcție de care să evalueze un produs *data mining* înainte de a se decide ce sistem *data mining* va achiziționa [26]. Dintre aceste criterii cele mai importante sunt:

- tipul de date de care compania dispune și pe care intenționează să le exploateze folosind tehnologii *data mining* (baze de date relaționale, text în format ASCII, pagini web, depozite de date etc.);
- surse de date (formatul specific al datelor pe care va lucra sistemul *data mining*);
- funcții și metodologii *data mining* disponibile. Este important să fie ales acel sistem *data mining* care are mai multe funcții și care să poată fi folosite cât mai direct pentru a oferi răspunsuri la problemele întreprinderii (clasificare, *clustering*, analiza valorilor aberante etc.). Un alt lucru important este ca fiecare funcție disponibilă să cuprindă cât mai multe metode pentru a oferi sistemului mai multă flexibilitate;
- gradul de cuplare a sistemului *data mining* cu un sistem de gestiune a bazelor de date sau cu depozite de date (este important, deoarece permite sistemului *data mining* să își optimizeze analiza);
- scalabilitatea (este foarte importantă pentru analiza unui volum de date mai mare decât cel prevăzut inițial);
- instrumente de vizualizare (oferă, pe lângă posibilitatea ca sistemul *data mining* să fie mai interactiv și mai prietenos, și o prezentare a rezultatelor de multe ori mai potrivită pentru a fi înțelese de către utilizatorul uman).

6.2. Caracteristici ale câtorva instrumente data mining disponibile pe piață

Sistemele *software* de *data mining* disponibile pe piață sunt de obicei produse ale unor companii din domeniul bazelor de date, hardware, din domeniul analizei statistice sau alte domenii apropiate. Unele dintre produsele cele mai cunoscute pe piață sunt prezentate în cele ce urmează, împreună cu caracteristicile principale ale fiecăruia, printre acestea numărându-se: SAS Enterprise Miner; IBM SPSS Modeler; instrumentele *data mining* incluse în Microsoft SQL Server - Data Mining – SSAS (Analysis Services); Oracle Data Miner, care lucrează cu baza de date de la versiunea Oracle 10g și mai noi; Angoss Knowledge STUDIO.

Dintre numeroasele produse *data mining* comerciale merită amintite produse realizate de firme cu tradiție în domeniul statisticii (cum ar fi, de exemplu, SPSS, SAS Enterprise Miner). De

asemenea, există produse ale unor companii din domeniul bazelor de date care în ultimii ani au produs şi instrumente *data mining*, dintre acestea cele mai reprezentative fiind SPSS Modeler, produs de firma IBM, *Microsoft SQL Server*, care deşi este un produs destinat în principal bazelor de date este completat şi de un set puternic de tehnici *data mining* - SSAS (Analysis Services) [39], *Oracle Data Mining* şi altele.

În afară de produsele realizate de firme cu tradiţie în domeniile statisticii sau ale bazelor de date, există şi produse ale altor firme cum ar fi Angoss Knowledge STUDIO, toate acestea având posibilitatea de a realiza extragerea de cunoştinţe şi informaţii folositoare din volume mari şi foarte mari de date.

O altă caracteristică importantă pe care o au produsele *data mining* comerciale este faptul că acestea au implementate un număr mare de tehnici şi funcţii *data mining*. Din acest motiv, putem considera că multe dintre acestea sunt suficient de generale pentru a se preta utilizării într-o gamă largă de domenii de activitate, acesta fiind un avantaj major care face ca aceste produse să fie atât de răspândite. De exemplu, produsul de *data mining Enterprise Miner,* creat de firma SAS, are un set bogat de algoritmi, dintre care merită să fie amintiţi cei de lucru cu arbori decizionali, algoritmii de lucru pe reţele neurale, algoritmii de regresie, precum şi algoritmii pentru îndeplinirea funcţiilor clasice de *data mining* cum ar fi asocierea, clasificarea, gruparea, analiza excepţiilor (*outlyer analysis*), având chiar şi posibilitatea de a face analize pe text nestructurat în baze de date sau depozite de date (lucru pe date în format neconvenţional).

Şi alte produse *data mining* prezintă diferite tipuri de tehnici şi funcţii *data mining*. De exemplu, produsele SPSS au posibilitatea de lucru cu arbori decizionali (*Answer Tree*), permit utilizatorilor să realizeze operaţii anterioare analizei propriu-zise cum ar fi curăţarea datelor în vederea păstrării a ceea ce este relevant din punctul de vedere al analizei şi transformarea datelor într-un format mai accesibil în care să fie exploatate.

Nici dezvoltarea produselor pentru *data mining* ale companiei Oracle nu este de neglijat. Oracle Data Mining a început cu versiunea Oracle şi cu algoritmi clasici de *data mining* precum asocierea, gruparea, clasificarea, Naive Bayes, pentru ca mai târziu, pentru versiunile mai noi de 10g, să se dezvolte foarte mult în diferite direcţii, la momentul actual cuprinzând un număr foarte mare de algoritmi îndeplinind diferite funcţii *data mining*.

Producătorii de sisteme *data mining* care au tradiţie în domeniul bazelor de date cum ar fi IBM, Microsoft sau Oracle au produs sisteme *data mining* pornind de la propriile baze de date, în timp ce alţi producători care erau oarecum independenţi de o anumită platformă de baze de date au pornit din start de la ideea de a crea instrumente *data mining* care să poată exploata mai multe surse de date (printre care şi baze de date de la diferiţi producători).

Multe sisteme pentru *data mining* au fost create şi pentru a fi compatibile cu diferite limbaje utilizate în domeniul bazelor de date, fiecare producător orientându-se spre limbaje care să fie

destul de general utilizate pentru a oferi flexibilitate sporită utilizatorilor finali, dar totodată mergând şi spre limbaje cu care au mai lucrat pentru propriile sisteme de baze de date. *Microsoft SQL Server,* de exemplu, implementează standardul OLE DB, folosind un limbaj specializat pentru *data mining* similar cu limbajul de interogare SQL [3]. Alte produse utilizează alte limbaje, de exemplu *Predictive Modelling Markup Language* de la IBM pentru exportarea modelelor de date găsite.

Anumite instrumente comerciale destinate analizei avansate a datelor prezintă şi interfeţe de tip API, astfel fiind Microsoft SQL Server sau *Oracle Data Mining* care integrează *Java Data Mining API.* Aceste interfeţe API conţin funcţii care pot fi apelate de către aplicaţiile clienţilor, fiind o facilitate în plus [11]. Folosind aceste funcţii, clienţii sunt capabili să integreze în propriile aplicaţii funcţii *data mining* fără să fie efectiv obligaţi să şi implementeze algoritmii *data mining* corespunzători.

Unele produse *data mining* pot realiza şi funcţii OLAP pe depozite de date, astfel încât datele multidimensionale pot să fie mult mai uşor de explorat.

Nu în ultimul rând, este important modul de vizualizare (prezentare) al rezultatelor funcţiilor de *data mining*. Trebuie să se aibă în vedere întotdeauna ca utilizatorii finali ai informaţiilor şi cunoştinţelor extrase din volume foarte mari de date în general nu sunt specialişti în statistică sau în informatică, aşa că modul de prezentare nu trebuie să fie neglijat.

În Tabelul 2 sunt prezentate mai concis aceste produse comerciale pentru *data mining* împreună cu cele mai reprezentative caracteristici ale acestora.

Tabelul 2

Principalele produse data mining disponibile pe piaţă [39]

Produs pentru *data mining*	**Caracteristici principale**
SAS Enterprise Miner	• Provine din domeniul statisticii; • Interfaţă grafică uşor de folosit; • Algoritmi specializaţi pentru data mining pe text - SAS® Text Miner; • Set bogat de algoritmi tradiţionali *data mining,* printre care se remarcă: arbori decizionali, reţele neurale, regresie asociere etc.
IBM SPSS Modeler	• Provine din domeniul statisticii/ bazelor de date; • Este folosit pentru a construi modele predictive şi pentru a efectua alte sarcini analitice; • Are o interfaţă vizuală care permite utilizatorilor să implementeze algoritmi statistici şi de extragere a datelor fără programare efectivă.
Microsoft SQL Server - Data	• Provine din domeniul bazelor de date;

Mining – SSAS (Analysis Services)	• Printre altele, prezintă algoritmi pentru arbori decizionali, predicție și grupare; • Mai mulți algoritmi personalizabili; • Surse multiple de date; • Prezintă un API pentru integrarea cât mai ușoară a facilităților *data mining* în aplicațiile utilizatorilor; • Oferă securitate bazată pe roluri prin SQL Server Analysis Services, inclusiv permisiuni separate pentru drillthrough pentru modelarea și structurarea datelor.
Oracle Data Miner	• Provine din domeniul bazelor de date; • A început cu algoritmi cum ar fi asocierea și *Naive Bayes* pentru ca după versiunea 10g să includă o foarte mare varietate de algoritmi; • Integrează *Java Data Mining API*, un pachet Java pentru includerea de facilități *data mining* în aplicațiile utilizatorului.
Angoss Knowledge STUDIO	• Prezintă algoritmi pentru construirea de arbori decizionali, analiză cluster (grupare), și modele predictive; • Permite utilizatorilor să exploateze datele sub diferite forme; • Prezintă instrumente puternice de vizualizare a rezultatelor, ceea ce îl face prietenos pentru utilizator; • Este compatibil cu alte baze de date, cum ar fi *Microsoft SQL Server*, putând interacționa cu acestea și la nivel de *data mining*.

6.3. Instrumente data mining utilizate în mediul de afaceri

Un studiu realizat de Daniel Andersson și Hannes Fries [7] încearcă să analizeze dacă întreprinderile folosesc instrumente *data mining* pentru a-și îmbunătăți capacitatea de decizie. Studiul este bazat pe un număr de 95 de companii mari din Suedia, alese la întâmplare, și are scopul de a vedea câte dintre acestea folosesc un sistem *data mining* și cum văd managerii acestora acest lucru. Rezultatele arată că mai puțin de 30% dintre întreprinderile analizate folosesc un sistem *data mining,* în timp ce mai mult de 50% dintre managerii întreprinderilor urmăresc cu interes evoluțiile în acest domeniu și intenționează să achiziționeze o astfel de soluție. Același studiu anticipează că piața sistemelor *data mining* va crește cu 100% în următorii patru ani. Astfel, se poate observa că întreprinderile devin mai încrezătoare în asemenea soluții pe măsură ce domeniul *data mining* se maturizează.

Rezultatele acestui studiu nu pot fi, desigur, extinse pentru alte țări fără a ține cont de diferențele de dezvoltare din punct de vedere tehnologic. Din această cauză, putem să presupunem

că în țări precum cele din Europa de est folosirea tehnologiilor *data mining* este mai redusă decât indică studiul amintit anterior, dar potențialul pieței ar putea fi totodată mai mare.

Ca o alternativă la utilizarea de produse comerciale ce presupun costuri mari de achiziție și implementare pe care firmele mici le pot suporta mai greu, este posibil ca acestea să-și implementeze proprii algoritmi **data mining** integrați cu aplicațiile pe care acestea le utilizează deja, adaptați la necesitățile fiecăreia. Acest fel de algoritmi ar avea avantajul că sunt gândiți pentru a rezolva doar niște probleme specifice, putând să aibă o complexitate mai redusă ca algoritmii comerciali, gândiți să acopere cât mai multe situații posibile. O asemenea abordare ar fi posibilă, în majoritatea cazurilor, pentru că majoritatea firmelor mici deja utilizează sisteme de gestiune a bazelor de date care pun la dispoziție și limbaje de programare.

7. CONFIDENȚIALITATEA DATELOR ANALIZATE CU INSTRUMENTE DATA MINING

Pe lângă numeroasele avantaje care decurg din folosirea sistemelor *data mining*, există totuși și anumite posibile probleme legate în special de confidențialitatea datelor analizate. „În urma analizei, nu cumva câștigăm mai puțin decât pierdem?" [14]. Atunci când căutăm cunoștințe în volume mari de date este foarte ușor a se scăpa din vedere problema confidențialității datelor, care ar putea fi compromisă, și acest aspect ar trebui analizat temeinic înainte de a lua hotărârea de a aplica tehnici *data mining*.

Putem lua ca exemplu două companii care ar putea să se alieze presupunând că fiecare deține ceva care celeilalte îi lipsește, dar rezultatul acestei decizii depinde de avantajele pe care această alianță le-ar aduce pentru fiecare din ele. Analizând datele de care dispun ambele companii folosind diverse tehnici *data mining,* s-ar putea ajunge la unele concluzii relevante pro sau contra deciziei de a se alia. Problema de confidențialitate apare în cazul în care rezultatele analizei ar indica faptul că alianța nu este potrivită. Dacă se consideră că datele fiecărei companii sunt confidențiale trebuie asigurată această confidențialitate și în timpul analizei.

Trebuie specificat faptul că, prin dezvoltarea pe care au căpătat-o, culegerea informațiilor și aplicarea tehnicilor *data mining* asupra acestora pot conduce la situații în care intimitatea anumitor persoane să fie amenințată. De exemplu, un magazin virtual poate stoca date despre comportamentul unui client pentru ca apoi să-i afișeze acele reclame care l-ar putea interesa, sau să-i scoată în față produsele care ar fi cel mai probabil să le cumpere, fără ca acesta să își dea seama că acțiunile lui sunt înregistrate pentru o analiză ulterioară.

Pentru a evita problemele legate de confidențialitatea datelor, este necesar ca o firmă care stochează date referitor la clienții săi să le ceară mai întâi acordul pentru a face acest lucru, sau cel puțin să-i

informeze de ceea ce se va întâmpla cu datele lor, cu respectarea legislației privind protecția datelor cu caracter personal în vigoare (GDPR).

Firmele care deţin baze de date cu informaţii confidenţiale despre clienţii acestora trebuie să-şi securizeze bazele de date şi să se asigure ca cei neautorizaţi nu vor avea acces la acestea.

Luând în considerare toate cele de mai sus, se poate afirma totuşi că majoritatea tehnicilor *data mining* nu aduc prejudicii informaţiilor personale ale celor implicaţi, ci, mai degrabă, acestea provin din securizarea deficitară a bazelor de date cu informaţii personale.

Tehnicile *data mining* încearcă, mai degrabă, să găsească modele general – valabile în datele existente decât să se concentreze pe detalii conţinute în volumul de date analizat. În plus, în multe dintre domeniile unde se aplică aceste tehnici, cum ar fi, de exemplu, biologia, geologia, geografia, astronomia sau alte ştiinţe nu se lucrează niciodată cu date personale ale cuiva, aşa că problema confidenţialităţii datelor analizate nu îşi are sensul.

CONCLUZII

Cu toate că tehnicile *data mining* au cunoscut o dezvoltare deosebită în ultimii ani şi s-a ajuns la o relativă maturitate a tehnologiei, însoţită de apariţia a tot mai numeroase instrumente comerciale dedicate, încă nu se poate spune că piaţa produselor *data mining* a ajuns la o dimensiune corespunzătoare. Acest lucru se datorează, în primul rând, faptului că produsele care implementează tehnologii relativ noi trebuie mai întâi să fie bine cunoscute de către potenţialii cumpărători pentru a putea fi comercializate cu succes. În cazul particular al produselor dedicate asistării managementului strategic al unei organizaţii economice, este nevoie de o bună informare a managerilor asupra acestor produse şi asupra avantajelor utilizării lor în obţinerea de avantaje competitive.

Instrumente specializate *data mining* au pătruns deja pe piaţă, bucurându-se de oarecare succes, în special, printre utilizatorii companiilor mari, deşi, cantitativ vorbind, se poate aprecia că şi această piaţă este încă în dezvoltare. Deşi avantajele utilizării unor astfel de produse sunt demonstrate şi evidente pentru marea majoritate a managerilor, majoritatea întreprinderilor mici le privesc cu reticenţă, de obicei, din cauza preţurilor încă destul de mari.

Pentru optimizarea proceselor economice care se desfăşoară în cadrul firmelor moderne, cu activitate complexă producţie-comerţ-servicii, se impune, pe de o parte, utilizarea unor instrumente informatice puternice şi performante, iar pe de altă parte, adoptarea unor măsuri bazate pe concluzii rezultate din aplicarea unor metodologii de optimizare recunoscute şi verificate în prealabil.

Tehnicile *data mining* pot fi aplicate pe date provenind din surse variate, dintre care cele mai uzuale sunt bazele de date şi depozitele de date. Pe lângă acestea au fost dezvoltate şi tehnici speciale *data mining* capabile să fie aplicate pe date provenind din fluxuri şi pe date puternic nestructurate cum ar fi datele în format multimedia, text nestructurat, pagini web etc.

Orice tehnică *data mining*, indiferent de sursa datelor pe care se aplică, trebuie să îndeplinească o funcţie *data mining*, dintre acestea cele mai comune fiind asocierea, clasificarea, analiza seriilor de timp etc. Aceste funcţii sunt realizate cu ajutorul unor algoritmi special proiectaţi cum ar fi Apriori pentru funcţia de asociere, *C4.5* pentru clasificare, *k-means* pentru grupare etc.

Au apărut produse comerciale pentru *data mining*, acestea fiind pachete software care implementează mai mulţi algoritmi *data mining* şi care sunt gândite să îndeplinească mai multe funcţii *data mining*. Aceste produse sunt de obicei costisitoare, în special pentru companiile mici, ceea ce face ca dezvoltarea pieţei să fie mult încetinită. În această lucrare, am căutat să identificăm soluţii mai puţin costisitoare pentru acestea.

BIBLIOGRAFIE

[1] Mihai Andronie, Modern data mining techniques, The Ninth International Conference on Informatics în Economy IE, Bucureşti 2009, p. 753-757.

[2] Puscasiu Adela; Fanca Alexandra; Gota Dan-Ioan; Valean Honoriu - PROCEEDINGS OF 2020 IEEE INTERNATIONAL CONFERENCE ON AUTOMATION, QUALITY AND TESTING, ROBOTICS (AQTR), Page 367-372, 2020.

[3] Mihai Andronie, Analiza datelor stocate în depozite mari de date, Sesiunea de comunicări ştiinţifice a cadrelor didactice din facultăţile economice ale Universităţii Spiru Haret, Bucureşti, 2008.

[4] Avram Anca; Matei Oliviu; Pintea Camelia-M; Pop Petrica C; Anton Carmen Ana - Context-Aware Data Mining vs Classical Data Mining: Case Study on Predicting Soil Moisture. 14TH INTERNATIONAL CONFERENCE ON SOFT COMPUTING MODELS IN INDUSTRIAL AND ENVIRONMENTAL APPLICATIONS (SOCO 2019).

[5] R. Agrawal, R. Srikant, Mining Sequential Patterns, The 11th International Conference on Data Engineering, 1995.

[6] Yu Bin; Mao Wenjie; Lv Yihan; Zhang Chen; Xie Yu. A survey on federated learning in data mining. WILEY INTERDISCIPLINARY REVIEWS-DATA MINING AND KNOWLEDGE DISCOVERY, Volume 12, Issue 1, 2021.

[7] Daniel Andersson, Hannes Fries, Data Mining Maturity. A Quantitative Study of Large Companies în Sweden, Jonkoping University, Master's Thesis în Informatics, 2008.

[8] Wu ChienHsing; Kuo ShangWei; Kao Shu-Chen. Classification-Based Data Mining Applied in Vehicle Accident Prediction. FUZZY SYSTEMS AND DATA MINING V (FSDM 2019), Volume 320, Page 218-223.

[9] Bartschat Andreas; Reischl Markus; Mikut Ralf. Data mining tools. WILEY INTERDISCIPLINARY REVIEWS-DATA MINING AND KNOWLEDGE DISCOVERY, Volume 9, Issue 4, 2019.

[10] Hassani Hossein; Beneki Christina; Unger Stephan; Mazinani Maedeh Taj; Yeganegi Mohammad Reza. Text Mining in Big Data Analytics. BIG DATA AND COGNITIVE COMPUTING, Volume 4, Issue 1, 2020.

[11] Berger Ch , Haberstroh B., Oracle Data Mining 10g Release 2, White Paper, 2005.

[12] B. Liu: Web Data Mining: Exploring Hyperlinks, Contents and Usage Data, Springer, 2007.

[13] C. Borgelt, Efficient Implementations of Apriori and Eclat, IEEE workshop Frequent Itemset Mining Implementations Proceedings, 2003.

[14] Chris Clifton, Don Marks, Security and Privacy Implications of Data Mining, ACM SIGMOD Workshop Proceedings, 1996.

[15] Collier K., Carey E, Grusy C., Marjaniemi D., Sautter D. – A Perspective on Data Mining, Northern Arizona University, 1998.

[16] B. Dasarathy, Nearest Neighbours (NN) Norms: NN Pattern Classification Techniques, IEEE Computer Society Press, 1991.

[17] Delobel C., Adiba M., Bases de donnees et systemes relationels, Dunod Informatique, 1982.

[18] Dicționarul explicativ al limbii române, ediția a II-a, Academia Română, Institutul de Lingvistică „Iorgu Iordan", Editura Univers Enciclopedic, 1998.

[19] R.O. Duda, P. E. Hart, D. G. Stork, Pattern Classification, Wiley, 2001.

[20] C. Elkan, Clustering with k-Means: Faster, Smarter, Cheaper, SIAM International Conference on Data Mining, 2004.

[21] Ingrid Fischer, Thorsten Meinl, Graph Based Molecular Data Mining – An Overview, 2005.

[22] W. Frawley, G. Piatetsky-Shapiro, C. Matheus, Knowledge Discovery în Databases: an Overview, AI Magazine, 1992.

[23] Y. Freund, R. Schapire, A Decision-Theoretic Generalisation of Online Learning and an Application to Boosting, 1997.

[24] D. Gondek, T. Hoffmann, Non Redundant Data Clustering, Knowledge Information Systems, 2007.

[25] Gupta, Bhatnagar, Wasan, Architecture for Knowledge Discovery and Knowledge Management. Knowledge and Information Systems, 2005.

[26] Jiawei Han, Micheline Kamber, Data Mining. Concepts and Techniques, Second Edition, Elsevier, 2006.

[27] David Hand, Heikki Mannila, Padhraic Smith, Principles of Data Mining, Prentice Hall India, 2006;

[28] J. Han, J. Pei, Y. Yin, Mining Frequent Patterns without Candidate Generation, Proceedings of ACM SIGMOD international Conference, 2000.

[29] Marti A. Hearst, Untangling Text Data Mining;

[30] Bidgoli, Hossein, The Internet Encyclopedia, John Wiley & Sons, 2004.

[31] A. Jamain, D. Hand, Mining Supervised Classification Performance Studies: A Meta-Analytic Investigation, Journal of Classification, 2008.

[32] Longbing, C. & Chengqi, Z., The Evolution of KDD: Towards Domain-driven Data Mining, International Journal of Pattern Recognition, 2007.

[33] Andrei Mihalache, Adrian Ghencea, Ion Lungu, Adaptive Information Systems – the premise for Self Learning Systems. Proceedings of the International Business Information Management Conference (13th IBIMA) on 9-10 November 2009, Marrakech, Morocco, ISBN: 978-0-9821489-2-1.

[34] V. Militaru, Studiu comparat asupra tehnicilor de data mining utilizate în rezolvarea problemelor de regresie şi clasificare, Informatica Economica, 2003.

[35] NASA: Workshop on Issues în the Application of Data Mining to Scientiffic Data, University of Alabama, 1999.

[36] The OLAP Council, OLAP and OLAP Server Definitions, 1995.

[37] A. Pavlo, E. Paulson, A. Rasin, D. Abadi, D. Dewitt, S. Madden, M. Stonebraker, A Comparisson of Approaches to Large-Scale Data Analysis, Brown University, 2010.

[38] A.Pirjan, The Optimization of Algorithms în the Process of Temporal Data Mining Using the Compute Unified Device Architecture, Database Systems Journal, No. 1/2010.

[39] Zhao Hui Tang, Jamie MacLennan, Data Mining with SQL Server 2005, Wiley, 2005.

[40] P. N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, Addison-Wesley, 2005.

[41] Velicanu Manole, Dicționar explicativ al sistemelor de baze de date, 2005.

[42] S. Wasserman, K. Raust, Social Network Analysys, Cambridge University Press, 1994.

[43] I. H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Second Edition, Morgan Kaufmann, 2005.

[44] Wu Xindong, Vipin Kumar, The Top Ten Algorithms în Data Mining, Chapman and Hall/CRC 2009.