UNIVERSITATEA BABEŞ-BOLYAI
FACULTATEA DE MATEMATICĂ ŞI INFORMATICĂ

Noi abordări în vederea extragerii cunoştinţelor din date. Aplicaţii în domeniul educaţional

Rezumatul tezei de doctorat

Student doctorand: Liana-Maria Crivei
Coordonator ştiinţific: Prof. Dr. Czibula Gabriela

Septembrie 2020

Cuvinte cheie: Extragerea cunoştinţelor din date de natură academică, predicţia performanţelor studenţilor, reguli de asociere relaţionale, ansambluri de arbori de decizie, reţele neuronale profunde, reţele cu autoorganizare.

Cuprins

Cuprinsul tezei

Lista publicaţiilor

Toate rangurile sunt listate conform clasificării revistelor [1] şi conferinţelor [2] din Informatică din 2014.

Publicaţii în Web of Science - Science Citation Index Expanded

[14] Czibula, G., Czibula, I.G., Miholca, D.L., **Crivei, M.L.**. A novel concurrent relational association rule mining approach. *Expert systems with Applications* 125(6), 2019, pp. 142–156. (**indexată Web of Science, FI=3.768**)

Rang A, 4 puncte.

Publicaţii în Web of Science - Conference Proceedings Citation Index

[25] Diana-Lucia Miholca, Gabriela Czibula, **Liana Maria Crivei**. *A new incremental relational association rules mining approach.* International Conference on Knowledge Based and Intelligent Information and Engineering Systems (KES), 2018, Procedia Computer Science, Volume 126, 2018, pp. 126–135. (**indexată Web of Science**)

Rang B, 4 puncte.

[15] Gabriela Czibula, Andrei Mihai, **Liana Maria Crivei**. *SPRAR: A novel relational association rule mining classification model applied for academic performance prediction.* International Conference on Knowledge Based and Intelligent Information and Engineering Systems (KES), Procedia Computer Science, Volume 159, 2019, pp. 20–29. (**indexată Web of Science**)

Rang B, 4 puncte.

[11] **Liana Maria Crivei**, Andrei Mihai, Gabriela Czibula. *A study on applying relational association rule mining based classification for predicting the academic performance of students.* The 12th International Conference on Knowledge Science, Engineering and Management (KSEM), LNAI 11775, 2019, pp. 287–300. (**indexată Web of Science**)

Rang B, 4 puncte.

[12] **Liana Maria Crivei**, Gabriela Czibula, George Ciubotariu, Mariana Dindelegan. *Unsupervised learning based mining of academic data sets for students' performance analysis.* IEEE 14th International Symposium on Applied Computational Intelligence and Informatics, SACI 2020, pp. 11–16. (**indexată Web of Science**)

[1] hfpop.ro/standarde/doctorat/2014-jurnale.pdf
[2] https://hfpop.ro/standarde/doctorat/2014-conferinte.pdf

Rang C, 1 punct.

Publicaţii în reviste şi proceedings de conferinţe indexate în baze de date internaţionale

[13] **Liana Maria Crivei**, Vlad-Sebastian Ionescu, Gabriela Czibula. An analysis of supervised learning methods for predicting students' performance in academic environments. *ICIC Express Letters*, 13(3), 2019, pp. 181–189 (**indexată Scopus**)

Rang C, 2 puncte.

[9] **Liana Maria Crivei**. Incremental relational association rule mining of educational data sets. *Studia Universitatis Babes-Bolyai Series Informatica*, 63(2), 2018, pp. 102–117. (**indexată Mathematical Reviews**)

Rang D, 1 punct.

[8] George Ciubotariu, **Liana-Maria Crivei**. *Analysing the academic performance of students using unsupervised data mining. Studia Universitatis Babes-Bolyai Series Informatica*, Vol. 64(2), 2019, pp. 34–48. (**indexată Mathematical Reviews**)

Rang D, 1 punct.

[10] **Liana-Maria Crivei**. *Using unsupervised learning models for analyzing students' academic performance.* PhD Coloquium at 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2018), 2018, pp. 1–4.

Scorul publicaţiilor: 21 puncte.

Introducere

Domeniul principal de cercetare al tezei de doctorat este extragerea de cunoştinţe din date de natură academică. *Extragerea de cunoştinţe din date de natura academică* (EDM) este un domeniu de cercetare în curs de dezvoltare care se concentrează pe analiza *mediilor educaţionale* din perspectiva *extragerii cunoştinţelor din date*. Scopul principal este de a înţelege mai bine fenomenele educaţionale şi de a descoperi, prin tehnici de extragere a datelor, şabloane ascunse relevante din datele educaţionale. Tehnicile de *extragere a cunoştinţelor din date* (DM) sunt aplicate în zilele noastre în diferite domenii, inclusiv medicină, bioinformatică, inginerie software, pentru a descoperi şabloane relevante în bazele de date mari, în special datorită potenţialului lor de a descoperi informaţii ascunse din date.

Aplicarea tehnicilor de *învăţare automată* în educaţie [1] a atras în mod continuu cercetătorii din domeniul *extragerii de cunoştinţe din date de natura academică*. Scopul principal în acest domeniu este acela de a descoperi şabloane relevante în date care provin din diverse medii educaţionale. Unul dintre scopurile *extragerii cunoştinţelor din date educaţionale* este de a oferi informaţii suplimentare asupra procesului de învăţare a studenţilor şi de a oferi astfel o mai bună înţelegere a fenomenelor educaţionale. Extragerea tiparelor relevante din procesele educaţionale ar fi de asemenea utilă pentru înţelegerea studenţilor şi a modului lor de învăţare, precum şi pentru îmbunătăţirea rezultatelor educaţionale (de exemplu, rezultatelor învăţării). Domeniul *extragerii cunoştinţelor din date educaţionale* a primit în ultima vreme o atenţie considerabilă din partea comunităţii de cercetare, deoarece extragerea cunoştinţelor ascunse din datele educaţionale este de interes deosebit pentru instituţiile academice şi este utilă şi pentru îmbunătăţirea metodologiilor lor de predare şi a proceselor de învăţare [29].

Extragerea *regulilor de asociere* (AR) reprezintă o tehnică importantă de analiză şi de extragere a datelor [7] utilă în procesele de *învăţare automată* pentru descoperirea unor tipare semnificative bazate pe reguli de asociere în seturile de date. *Regulile de asociere ordinale* (OAR) [6] au fost propuse ca o clasă particulară a *regulilor de asociere* care exprimă relaţii ordinale între atributele care caracterizează un set de date. *Regulile de asociere relaţională* (RAR) [4, 3] au fost introduse ca o extensie a *regulilor de asociere ordinale*, fiind capabile să exprime diferite tipuri de relaţii non-ordinale între atributele datelor. Atât extragerea regulilor de asociere ordinală, cât şi relaţională, au fost aplicate cu succes pentru a rezolva probleme care variază de la sarcinile de extragere a datelor (de exemplu, detectarea defectelor de proiectare software [18] sau curăţarea datelor [4]) până la clasificarea supervizată (de exemplu, predicţia defectelor software [26]).

Motivaţie

Datorită capacităţii lor de a exprima relaţii relevante în date, aplicarea regulilor de asociere relaţionale (RAR) pe seturi de date academice ar putea oferi abordări noi şi eficiente pentru rezolvarea diverselor sarcini EDM. Un exemplu concret în care RAR-urile

ar fi mult mai semnificative decât AR-urile clasice este legat de domeniul *extragerii de cunoştinţe din date de natura academică* (EDM) care se referă la seturile de date din mediile educaţionale. Într-un scenariu în care setul de date care trebuie extras conţine informaţii privind performanţa academică a studenţilor (de exemplu, fiecare student este caracterizat prin notele sale), RAR-urile ar putea să exprime relaţii relevante între notele studenţilor. Astfel, RAR-urile pot oferi informaţii suplimentare în ceea ce priveste predicţia performanţei studenţilor.

Literatura EDM dezvăluie că DM este foarte utilă în domeniul educaţional, în special atunci când se explorează mediul de învăţare online [29]. În acest sens, datorită relevanţei sale practice pentru extragerea de date *online*, problema *extragerii incrementale a regulilor de asociere relaţională* (IRARM) în contextul unui set de date dinamic, la care se adaugă noi instanţe este atractivă pentru cercetătorii DM. Întrucât procesele de învăţare în mediile educaţionale sunt prin natura lor procese online, ideea investigării perspectivei RAR incrementale asupra EDM apare în mod natural.

Datorită NP-completitudinii sale, problema extragerii tuturor *regulilor de asociere relaţională* interesante dintr-un set de date este dificilă din punct de vedere computaţional. Pe măsură ce dimensionalitatea setului de date analizat creşte, algoritmul clasic de *Descoperire a regulilor de asociere relaţională* (*DRAR*) pentru extragerea RAR-urilor interesante eşuează să furnizeze setul de reguli într-un timp rezonabil. În scopul reducerii duratei generale de extragere, vom investiga o perspectivă de extragere a *regulilor de asociere relaţionale concurente* folosind concurenţa pentru procesul de descoperire a RAR-urilor.

În cadrul cercetării EDM, dezvoltarea unor modele de învăţare automată pentru analiza şi predicţia performanţei studenţilor este de mare interes. Ca etapă naturală anterioară dezvoltării de noi modele de învăţare automată pentru predicţia performanţei studenţilor, este investigată utilitatea aplicării unor *metode de învăţare automată nesupervizată* în analiza datelor de performanţă academică ale studenţilor. Ca pas ulterior, datorită capacităţii lor de a descoperi cunoştinţe semnificative din date, RAR-urile vor fi explorate în contextul prezicerii performanţei academice a studenţilor. Mai precis, sunt avute în vedere *modele de învăţare automată* supervizată bazate pe reguli de asociere relaţională.

Pentru a rezuma, cercetarea realizată în teza de doctorat va fi concentrată pe trei direcţii principale: (1) contribuţii conceptuale în domeniul extragerii RAR prin introducerea de noi abordări pentru extragerea *incrementală* şi *concurentă* RAR; (2) cercetarea şi evidenţierea relevanţei şi eficienţei extragerii *regulilor de asociere relaţională* din seturi de date educaţionale; şi (3) dezvoltarea modelelor de *învăţare automată* supervizate şi nesupervizate pentru predicţia performanţei academice a studenţilor bazată pe RAR-uri şi sisteme bazate pe reguli. Mai precis, obiectivele majore ale cercetării realizate în această teză sunt: utilizarea extragerii regulilor de asociere relaţională în domeniul EDM; dezvoltarea de noi modele de învăţare automată (bazate în principal pe extragerea RAR) pentru a prezice performanţa academică a studenţilor; dezvoltarea unei versiuni concurente a algoritmului de extragere a RAR pentru a reduce timpul de calcul al procesului de extragere şi investigarea unei versiuni incrementale a algoritmului de extragere RAR, care ar fi adecvată în scenarii de extragere online în care sunt adăugate noi instanţe la setul de date care urmează să fie extras.

Structura tezei

Restul tezei este structurat după cum urmează.

Capitolul 1 prezintă principalele concepte privind EDM şi extragerea regulilor de asociere,

împreună cu problemele pe care intenţionăm să le abordăm din perspectivă computaţională. Importanţa EDM în contextul învăţării automate este subliniată în secţiunea 1.1. Abordăm concepte de învăţare supervizată şi nesupervizată şi prezentăm o imagine cuprinzătoare a literaturii asupra lucrărilor de cercetare existente în acest domeniu. De asemenea, includem o prezentare a literaturii despre predicţia performanţei instructorilor şi alte studii EDM. Secţiunea 1.2 prezintă conceptele fundamentale de seturi frecvente şi reguli de asociere clasice, precum şi o descriere cuprinzătoare a regulilor de asociere relaţionale, definiţia acestora şi un exemplu practic. În această secţiune este de asemenea prezentată o imagine de ansamblu a literaturii existente privind extragerea regulilor de asociere incrementale şi concurente. Ultima parte a acestui capitol, secţiunea 1.3 introduce trei modele de învăţare automată fundamentale: reţele neuronale artificiale, păduri aleatoare de arbori şi reţele cu auto-organizare, care vor fi ulterior folosite pentru predicţia performanţei studenţilor.

Capitolul 2 prezintă cercetarea originală asupra extragerii incrementale a regulilor de asociere relaţională în contextul seturilor de date dinamice, precum şi extragerea concurentă a RAR. În sectiunea 2.1 introducem o nouă metodă incrementală numită *extragerea incrementala a regulilor de asociere relationala* ($IRARM$). Prin $IRARM$ intenţionăm să ajustăm progresiv RAR-urile interesante descoperite într-un set de date original care este ulterior extins cu noi instanţe primite. Obiectivul nostru este obţinerea unui timp de calcul redus în comparaţie cu cel obţinut prin intermediul metodei $DRAR$ aplicate de la zero pe setul de date extins. Mai mult, este prezentată metodologia utilizată pentru $IRARM$, cu algoritmul de calcul şi o defalcare a celor două etape numite $filtering$ şi $extending$. Sunt efectuate experimentele pe două studii de caz pentru a evalua şi evidenţia eficienţa timpului de extragere $IRARM$. Secţiunea 2.2 introduce o nouă abordare de extragere concurentă numită $CRAR$ *extragerea concurentă a regulilor de asociere relaţională*. Noua metodă foloseşte concurenţa pentru descoperirea RAR-urilor şi scade considerabil timpul de extragere. Performanţa $CRAR$ din punct de vedere al timpului de calcul este evaluată şi comparată cu algoritmul secvenţial $DRAR$. O prezentare exhaustivă a algoritmului $CRAR$ este inclusă împreună cu funcţiile pentru generarea candidaţilor (*candidate generation*), partiţionare (*partitioning*) şi generarea regulilor (*rule generation*) pe mai multe fire de execuţie. Evaluările experimentale sunt efectuate pe nouă seturi de date disponibile public din diverse domenii: inginerie software, extragere de date educaţionale şi bioinformatică. Implementarea $CRAR$ pe seturile de date menţionate anterior demonstrează că timpul de rulare al algoritmului $CRAR$ este semnificativ mai mic în comparaţie cu metoda $DRAR$ aplicată de la zero.

Capitolul 3 prezintă contribuţiile şi rezultatele originale privind dezvoltarea modelelor de învăţare nesupervizate în contextul EDM. Secţiunea 3.1 explorează două modele de învăţare automată nesupervizată, în special *reţele cu auto-organizare* şi *extragerea regulilor de asociere relaţională* pentru a evalua performanţa studenţilor în medii academice. Analiza se realizează pe un set de date academice reale în care atributele fiecărei instanţe reprezintă performanţa studenţilor la un anumit curs academic, de-a lungul unui semestru. Studiul nostru dezvăluie potenţialul celor două tehnici de învăţare nesupervizată de a identifica şabloane relevante şi relaţii în cadrul datelor academice care sunt valoroase pentru predicţia realizărilor academice ale studenţilor. În secţiunea 3.2 extragerea RAR clasică şi metoda $IRARM$ anterior introduse în secţiunea 2.1 sunt testate şi evaluate în contextul datelor educaţionale. Experimentele sunt efectuate având în vedere două studii de caz şi un total de patru seturi de date academice. Studiul subliniază importanţa şi eficacitatea extragerii RAR (atât incrementale, cât şi clasice) pentru descoperirea cunoştinţelor şi şabloanelor semnificative din datele educaţionale.

Capitolul 4 descrie contribuţiile noastre în dezvoltarea modelelor de învăţare supervizată

pentru predicţia performanţei studenţilor. Secţiunea 4.1 analizează metode supervizate utile pentru predicţia performanţei studenţilor în medii academice. Două modele de regresie, *păduri aleatoare de arbori* (RF) şi *reţele neuronale artificiale* (ANN), sunt cercetate din punct de vedere al performanţei predicţiei. Experimentele sunt efectuate pe trei seturi de date academice reale care conţin notele obţinute de studenţi în timpul semestrului la diferite cursuri la nivel de licenţă. Performanţa predicţiei este măsurată folosind *rădăcina pătrată a erorii medii pătratice* ($RMSE$) şi măsura F - *score*. Este prezentată, de asemenea, o discuţie cu privire la rezultatele obţinute, precum şi o comparaţie cu alte lucrări din domeniu. Secţiunea 4.2 introduce un nou model de clasificare, $SPRAR$ (*predicţia performanţei studenţilor folosind reguli de asociere relaţională*) pentru predicţia rezultatului final al unui student la o anumită disciplină academică (pe baza notelor primite în timpul semestrului) folosind *reguli de asociere relaţională* (RAR). Experimentele sunt efectuate utilizând trei seturi reale de date academice. Performanţa clasificatorului $SPRAR$ pe studiile de caz considerate este comparată cu lucrări similare existente, fiind superioară rezultatelor obţinute de abordări propuse în literatură. Un studiu comparativ asupra modelului de clasificare $SPRAR$ introdus anterior este efectuat în continuare în secţiunea 4.3. În această secţiune sunt introduse trei noi scoruri de clasificare, care sunt utilizate în etapa de clasificare a $SPRAR$. Performanţa lor este analizată pe aceleaşi seturi de date academice reale utilizate în cercetarea iniţială din secţiunea 4.2 şi este comparată cu rezultatele similare existente în literatura.

Contribuţii originale

Teza are trei contribuţii majore, care sunt cuprinse în Capitolele 2, 3 şi 4 după cum urmează:

1. Contribuţii conceptuale în domeniul DM, în special în direcţia exragerii *incrementale* şi *concurente* a RAR, introduse în capitolul 2 şi publicate în lucrările originale [25] şi [14].

    - O metodă de extragere incrementală, numită $IRARM$ *extragerea incrementală a regulilor de asociere relaţională*, este introdusă în secţiunea 2.1 pentru a extrage eficient toate RAR-urile interesante dintr-un set de date care a fost mărit (adică s-au adăugat noi entităţi) [25]. $IRARM$ adaptează setul tuturor RAR-urilor interesante descoperite anterior în setul de date iniţiale pentru a calcula noul set al tuturor RAR-urilor care sunt interesante în setul de date extins. Algoritmul $IRARM$ este evaluat experimental pe două seturi de date disponibile public. Experimentele efectuate subliniază faptul că $IRARM$ este capabil să returneze setul complet de RAR-uri interesante mai eficient decât în cazul aplicării algoritmului de extragere de la zero pe setul de date extins. Lucrarea originală a fost publicată în [25].
    - În secţiunea 2.2 este introdusă o noua abordare denumita $CRAR$ care presupune extragerea concurentă de *reguli de asociere relaţionale*. Eficacitatea extragerii este validată empiric pe nouă seturi de date disponibile public. Reducerea timpului de extragere atunci când se utilizează $CRAR$ comparativ cu algoritmul secvenţial $DRAR$ subliniază faptul că acesta poate fi aplicat cu succes în diferite scenarii practice de extragere a datelor. Lucrarea originală a fost publicată în [14].

2. Noi modele de învăţare nesupervizată în domeniul EDM, mai precis pentru *predicţia*

performanţei studenţilor (SPP) sunt introduse în capitolul 3 şi au fost publicate în lucrările originale [9], [8] şi [10].

- Modele de *învăţare nesupervizată* pentru analiza datelor de performanţa academica a studenţilor sunt propuse în secţiunea 3.1. Experimentele efectuate utilizând un set de date academice reale au evidenţiat potenţialul modelelor de *învăţare nesupervizată* pentru descoperirea cunoştinţelor semnificative în cadrul datelor educaţionale. Lucrările originale a fost publicate în [10] şi [8].
- Am investigat şi am evidenţiat relevanţa şi eficacitatea *extragerii regulilor de asociere relaţională* din seturile de date educaţionale în secţiunea 3.2. *Extragerea incrementală a regulilor de asociere relaţională* folosind abordarea $IRARM$ introdusă anterior în Capitolul 2 este aplicată în contextul EDM, subliniind relevanţa şi eficacitatea ei. Lucrarea originală a fost publicată în [9].

3. Noi *modele de învăţare automată* supervizată pentru SPP, construite pe sisteme bazate pe reguli şi RAR-uri sunt dezvoltate în Capitolul 4 şi au fost publicate în lucrările originale [13], [15] şi [11].

   - În secţiunea 4.1 realizăm un studiu asupra aplicării a două modele de regresie, *păduri aleataore de arbori* (RF) şi *reţele neuronale artificiale* (ANN), pentru predicţia performanţei academice a studenţilor. Pentru fiecare regresor, două modele de calcul sunt introduse pentru predicţia notei finale la examen pentru un student, bazată pe notele sale primite în timpul unui semestru academic. Lucrarea originală a fost publicată în [13].
   - În secţiunea 4.2 introducem un nou model de clasificare, $SPRAR$ (*predicţia performanţei studenţilor folosind reguli de asociere relaţională*) pentru predicţia, pe baza notelor primite în timpul semestrului, rezultatului final al unui student la o anume disciplină academică folosind *regulile de asociere relaţională* (RAR). Experimentele sunt efectuate pe trei seturi de date academice reale colectate de la Universitatea Babeş-Bolyai din România. Performanţa clasificatorului $SPRAR$ pe studiile de caz considerate este comparată cu cea a clasificatorilor existenţi pentru predicţia performanţei studenţilor, fiind superioară acestora. Lucrarea originală a fost publicată în [15].
   - Un studiu comparativ asupra modelului de clasificare introdus anterior $SPRAR$ este efectuat în continuare în secţiunea 4.3. Trei noi scoruri de clasificare sunt introduse în această secţiune şi sunt utilizate în etapa de clasificare a $SPRAR$. Performanţa lor este analizată pe aceleaşi seturi reale de date academice utilizate în lucrarea originală [15] şi comparată cu rezultatele similare existente în literatura de specialitate. Lucrarea originală a fost publicată în [11].

Capitolul 1

Concepte introductive

În acest capitol prezentăm conceptele fundamentale cunoscute legate de domeniul *extragerii de cunoştinţe din date de natură academică* (EDM), precum şi principalele concepte legate de modelele de calcul pe care le vizăm în cercetarea noastră.

1.1 Extragerea de cunoştinţe din date de natură academică (EDM)

Tehnicile *extragerii de cunoştinţe din date* (DM) sunt des aplicate în prezent în diverse domenii, inclusiv medicină, inginerie software, bioinformatică, pentru a descoperi şabloane relevante în bazele de date mari, în special datorită potenţialului acestora de a descoperi informaţii ascunse în seturile de date.

Aplicarea tehnicilor DM în educaţie [1] a atras cercetători atât din DM cât şi din cercetarea educaţională şi astfel a apărut o nouă disciplină de cercetare interdisciplinară cunoscută sub numele de *extragerea de cunoştinţe din date de natură academică* (EDM). Principalul accent în EDM este de a dezvolta metode de extragere a cunoştinţelor din date care provin din diverse sisteme de informaţii şi medii educaţionale. Prin extragerea de cunoştinţe din seturile de date educaţionale, scopul EDM este de a înţelege mai bine procesul de învăţare al studenţilor şi, astfel, de a oferi informaţii suplimentare asupra fenomenelor legate de educaţie.

EDM este un domeniu de cercetare în dezvoltare în care conceptul care stă la baza lui este acela de a aduce perspectiva *extragerii de cunoştinţe din date* în *medii educaţionale*. Scopul principal este de a înţelege mai bine fenomenele legate de educaţie prin extragerea tiparelor ascunse semnificative din datele educaţionale folosind tehnici de extragere a datelor.

Extragerea şabloanelor relevante din procesele educaţionale ar fi de asemenea utilă pentru înţelegerea studenţilor şi a modului în care învaţă, precum şi îmbunătăţirea rezultatelor educaţionale (de exemplu, rezultatele învăţării). EDM a primit în ultimul timp o atenţie considerabilă din partea comunităţii de cercetare, deoarece extragerea cunoştinţelor ascunse din datele educaţionale este de interes deosebit pentru instituţiile academice şi este de asemenea utilă pentru îmbunătăţirea metodologiilor de predare şi a proceselor de învăţare [29].

1.2 Extragerea regulilor de asociere din date

În această secţiune prezentăm conceptele fundamentale legate de extragerea *regulilor de asociere relaţională* (RAR) [3]. Extragerea *regulilor de asociere* (AR) reprezintă o importantă metodă de analiză a datelor şi tehnică de extragere a cunoştinţelor [7] aplicată în

diferite sarcini legate de *învăţarea automată* pentru descoperirea modelelor relevante bazate pe reguli în seturile de date. *Regulile de asociere ordinală* (OAR) [6] au fost propuse ca o clasă particulară de AR care exprimă relaţii ordinale între atributele care caracterizează un set de date. *Regulile de asociere relaţională* (RAR) [5, 3] au fost introduse ca o extensie a OAR-urilor capabile să exprime diferite tipuri de relaţiile neordinale între atributele datelor.

Regulile de asociere (AR) clasice [33] nu iau în considerare relaţiile care pot apărea între valorile atributelor, ci numai co-ocurenţa lor. Spre deosebire de acestea, *regulile de asociere ordinală* (OAR) [24] exprimă relaţii ordinale care apar frecvent în date. Dar pot exista şi relaţii informative între valorile atributelor care nu sunt ordinale.

Noţiunea de *regulă de asociere relaţională* (*RAR*) este definită în următoarele paragrafe.

Considerăm o mulţime $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ de *instanţe* sau *înregistrări*. Fie $\Omega = (a_1, \ldots, a_m)$ un şir de m atribute care caracterizează fiecare instanţă din setul de date $\mathcal{D}$. Fiecare atribut a_i ia valori într-un domeniu nevid şi non-fuzzy Δ_i, care conţine şi o valoare *nulă* (*vidă*). Notăm cu $\Psi(d_j, a_i)$ valoarea atributului a_i pentru o instanţă d_j.

Notăm cu $\mathcal{T}$ mulţimea tuturor relaţiilor posibile (care nu sunt neapărat ordinale) care pot fi definite între două domenii Δ_i şi Δ_j.

Definition 1. O *regulă de asociere relaţională [3]* este o expresie $(a_{i_1}, a_{i_2}, a_{i_3}, \ldots, a_{i_h}) \Rightarrow (a_{i_1}\tau_1 a_{i_2}\tau_2 a_{i_3} \ldots \tau_{h-1} a_{i_h})$, unde $\{a_{i_1}, a_{i_2}, a_{i_3}, \ldots, a_{i_h}\} \subseteq \Omega$, $a_{i_k} \neq a_{i_p}$, $k, p = 1, \ldots, h$, $k \neq p$ şi $\tau_k \in \mathcal{T}$ este o relaţie pe $\Delta_{i_k} \times \Delta_{i_{k+1}}$, Δ_{i_k} fiind domeniul atributului a_{i_k}.

a) Dacă $a_{i_1}, a_{i_2}, \ldots, a_{i_h}$ nu lipsesc din m instanţe din setul de date atunci numim $s = \frac{m}{n}$ *suportul* regulii.

b) Dacă notăm cu $D' \subseteq \mathcal{D}$ setul de instanţe unde $a_{i_1}, a_{i_2}, a_{i_3}, \ldots, a_{i_h}$ nu lipsesc şi toate relaţiile

$$\Psi(d_j, a_{i_1})\tau_1\Psi(d_j, a_{i_2}), \Psi(d_j, a_{i_2})\tau_2\Psi(d_j, a_{i_3}), \ldots, \Psi(d_j, a_{i_{h-1}})\tau_{h-1}\Psi(d_j, a_{i_h})$$

au loc pentru fiecare instanţă d din D' atunci numim $c = \frac{|D'|}{n}$ confidenţa regulii.

RAR-urile *interesante* au fost definite în [3] ca acele reguli care au atât suportul cât şi confidenţa mai mari sau egale cu un prag minim specificat. Pentru extragerea RAR-urilor interesante, un algoritm de tip Apriori, numit *DRAR* (descoperirea regulilor de asociere relaţională) a fost propus în [16] ca o extensie a algoritmului *DOAR* introdus în [6] pentru descoperirea OAR-urilor.

Algoritmul *DRAR* constă în generarea iterativă a RAR-urilor, începând cu regulile de lungime 2, apoi continuând cu regulile de lungime 3, ş.a.m.d. Setul de reguli de lungime 2 este filtrat pentru a determina regulile interesante (adică acele reguli care verifică suportul minim şi cerinţele legate de confidenţa). La un pas dat, pentru determinarea RAR-urilor de o anumită lungime l, începem de la setul de RAR-uri interesante de lungime $l-1$ generate la pasul anterior. Acest set este folosit pentru a genera prin alăturare noi posibile RAR-uri interesante, numite reguli candidat, care vor fi filtrate pentru a păstra doar regulile care sunt interesante. După ce se generează RAR-uri interesante de lungime l, procesul iterativ continuă cu generarea regulilor de lungime $l+1$. Procesul se opreşte atunci când nu mai sunt descoperite noi RAR-uri interesante.

1.3 Modelele de învăţare automată folosite

Această secţiune prezintă modelele de învăţare supervizată şi nesupervizată utilizate în teză pentru predicţia performanţei academice a studenţilor: *reţele neuronale artificiale*, *păduri aleatoare de arbori* şi *reţele cu auto-organizare*.

1.3.1 Reţele neuronale artificiale

Reţelele neuronale artificiale sunt utilizate pe scară largă ca modele de învăţare supervizată pentru diverse aplicaţii cum ar fi recunoaşterea tiparelor din date, recunoaşterea vorbirii [31], predicţia [23], identificarea şi controlul sistemelor. În mod similar sistemelor neuronale biologice, reţelele neuronale artificiale [28] constau dintr-un set de unităţi de calcul dens interconectate, numite *neuroni*.

O *reţea neuronală artificială* (ANN) [35] este un sistem adaptiv care învaţă o corespondenţă (o funcţie de intrare/ieşire) din date, prin reglarea autonomă a parametrilor sistemului în timpul *fazei de antrenament*. Parametrii reţelei obţinuţi după finalizarea antrenamentului sunt folosiţi în continuare pentru a rezolva problema curentă (*faza de testare*).

Într-un scenariu de învăţare supervizată, o instanţă de intrare este prezentată reţelei neuronale artificiale împreună cu ţinta corespunzătoare de ieşire [27]. Aceste exemple de intrare-ieşire sunt adesea furnizate de un supervizor extern. O eroare este reprezentată prin diferenţa dintre ieşirea dorită şi ieşirea sistemului. Această informaţie de eroare este propagată înapoi în reţea şi parametrii sistemului sunt ajustaţi prin *regula de învăţare*. Reţeaua neuronală artificială este construită folosind o procedură iterativă al cărei obiectiv este de a reduce la minimum eroarea obţinută, care este denumită în mod obişnuit *regulă de învăţare*. Acest proces se repetă până când se obţine o performanţă acceptabilă. Algoritmul pentru construirea modelului ANN se numeşte algoritmul de *propagare înapoi* (*backpropagation*).

1.3.2 Păduri aleatoare de arbori

Pădurile aleatoare de arbori (RF) [2] sunt o metodă de învăţare bazată pe ansambluri constând din agregarea mai multor predictori de tip arbore de decizie. În timpul procesului de construcţie al fiecărui arbore individual, se consideră pentru analiză numai o submulţime aleatoare de caracteristici şi o submulţime aleatoare de exemple de instruire. În acest fel, se evită supraantrenarea şi se obţine o stabilitate mai bună pentru generalizare. Eroarea de generalizare a pădurilor converge la limită, deoarece numărul arborilor din pădure devine mare, datorită legii numerelor mari. Această eroare depinde de puterea arborilor individuali din pădure şi de corelaţia dintre ei. Folosind o selecţie aleatoare de funcţii pentru a împărţi fiecare nod se produc rate de eroare care se compară favorabil cu algoritmul Adaboost [34], dar sunt mai robuste în ceea ce priveşte zgomotul. Fiind construite pe arbori de decizie, pădurile aleatoare de arbori pot fi utilizate în probleme de clasificare şi regresie.

1.3.3 Reţele cu auto-organizare

Reţelele cu auto-organizare (SOM) sunt modele de învăţare nesupervizată, de asemenea cunoscute în literatura de specialitate sub numele de *hărţi topografice Kohonen*, au fost introduse de Teuvo Kohonen şi sunt utilizate pe scară largă ca instrumente pentru vizualizarea datelor de dimensiuni mari. SOM-urile sunt legate de *reţelele neuronale artificiale* (ANN) din literatură şi de *învăţarea competitivă*. În *învăţarea competitivă*, neuronii de ieşire concurează pentru a fi activaţi. O *reţea cu auto-organizare* [32] este instruită folosind un algoritm de

învăţare nesupervizată (algoritmul lui Kohonen) pentru a realiza o corespondenţă, folosind o asociere neliniară, între spaţiul de intrare continuu format din instanţe multidimensionale şi un spaţiu de ieşire discret (de obicei bidimensional) numit *hartă* [19]. *Corespondenţa care păstrează topologia* este principala caracteristică oferită de un SOM. Aceasta înseamnă că instanţele de intrare care sunt aproape una de cealaltă în spaţiul de intrare vor fi, de asemenea, aproape una de cealaltă pe hartă (spaţiul de ieşire). Astfel, un SOM este capabil să furnizeze grupe de instanţe de date similare [22].

Reţelele cu auto-organizare sunt o clasă de reţele neuronale artificiale care utilizează o funcţie de vecinătate pentru a menţine caracteristicile topologice ale spaţiului de intrare. SOM-urile folosesc învăţarea nesupervizată pentru antrenament şi ele sunt unul dintre cele mai inovatoare tipuri de reţele neuronale artificiale din literatura de specialitate actuală. Algoritmul SOM foloseşte nodurile (neuronii) din reţea în manieră nesupervizată şi competitivă, neuronii concurând între ei pentru a răspunde la un subset de date primite. Reţeaua cu auto-organizare păstrează topologia spaţiului de intrare, care este pus în corepondenţă cu unul de dimensiune mică, de obicei un spaţiu bidimensional, denumit hartă. Reducerea dimensionalităţii prin păstrarea topologiei înseamnă că punctele de date de intrare vecine sau similare vor fi transformate în unităţi apropiate in spaţiul de ieşire [21]. Prin urmare, SOM-urile pot fi privite ca un instrument de vizualizare a datelor de dimensiune mare, precum şi ca un instrument de grupare.

Capitolul 2

Noi abordări în extragerea cunoştinţelor din date

Acest capitol introduce contribuţiile tezei noastre de doctorat în domeniul *extragerii de cunoştinţe din date*, mai precis în dezvoltarea de noi abordari privind extragerea *incrementală* şi *concurentă* a regulilor de asociere relaţională.

2.1 O abordare privind extragerea incrementală a regulilor de asociere relaţională

În primul rând, o nouă metodă incrementală numită $IRARM$ (*extragerea incrementală a regulilor de asociere relaţională*) este introdusă şi evaluată în termeni de eficienţă a timpului de execuţie. $IRARM$ îşi propune să identifice RAR-urile interesante dintr-un set de date de natură dinamică, ceea ce înseamnă că noi instanţe sunt adăugate treptat la setul de date iniţial. Caracteristic pentru $IRARM$ este că adaptează RAR-urile relevante descoperite anterior fără a necesita o rescanare a întregului set de date actualizat, spre deosebire de metoda $DRAR$. Eficacitatea $IRARM$ este evaluată prin experimente efectuate pe două seturi de date disponibile public, iar rezultatele evidenţiază capacitatea $IRARM$ de a îmbunătăţi semnificativ timpul de extragere comparativ cu $DRAR$.

Pe viitor, vor fi luate în considerare mai multe seturi de date, pentru a extinde şi mai mult evaluarea $IRARM$. De asemenea, ne propunem să investigăm condiţiile exacte în care este mai eficient să se aplice metoda incrementală $IRARM$ decât să se ruleze de la zero algoritmul offline $DRAR$. Mai mult, ne propunem să utilizăm IRARM în sarcini de extragere a datelor dinamice concrete, cum ar fi *predicţia defectelor software* incrementală şi extragerea incrementală de cunoştinţe din date educaţionale (*extragerea cunoştinţelor din date educaţionale* [30]). În plus, intenţionăm să investigăm o abordare incrementală adaptivă pentru extragerea *regulilor de asociere relaţională*, în care atât funcţii noi, cât şi obiecte noi sunt adăugate la setul de date extrase. Pentru îmbunătăţirea eficienţei procesului de extragere, va fi de asemenea avută în vedere o abordare distribuită.

2.2 O abordare privind extragerea concurentă a regulilor de asociere relaţională

A doua secţiune a acestui capitol introduce o nouă abordare privind extragerea concurentă a RAR, numită *CRAR* (*extragerea concurentă a regulilor de asociere relaţională*). Programarea concurentă este frecvent utilizată pentru a îmbunătăţi performanţa, exploatarea resurselor şi, de asemenea, pentru a dezvolta sisteme tolerante la erori şi scalabile. Algoritmul *CRAR* a fost dezvoltat pe aceste principii cu scopul de a furniza un timp de extragere mai bun în procesul de descoperire RAR. Efectuăm o evaluare experimentală detaliată a *CRAR*, realizând experimente pe nouă seturi de date disponibile public din diverse domenii: extragerea datelor educaţionale, predicţia defectelor software, bioinformatică. *CRAR* oferă o valoare semnificativ îmbunătăţită a timpului de exceuţie, prin reducerea timpului de extragere în medie cu 52,3 % în comparaţie cu algoritmul clasic *DRAR*. Abordările prezentate în capitolul curent sunt lucrări de cercetare originale publicate în [25] şi [14].

Pe viitor, ne propunem să investigăm optimizări ale *CRAR*, cum ar fi creşterea volumului de calcule efectuate în paralel, în special calculul RAR-urilor binare, care ar creşte *eficienţa scalării* şi *accelerarea paralelizării* algoritmilor. Pentru a evidenţia mai bine utilitatea *CRAR*, intenţionăm să utilizăm extragerea regulilor de asociere relaţională pentru probleme concrete de clasificare supervizată din *inginerie software bazată pe căutare* (de ex. *predicţia defectelor software* [27]) şi *extragerea de cunoştinţe din date educaţionale* [9]) (cum ar fi *predicţia performanţei studenţilor*). Cercetări suplimentare vor fi, de asemenea, axate pe extinderea abordării introduse de *CRAR* în această lucrare pentru extragerea *regulilor de asociere relaţională graduală* [17] în loc de RAR-urile clasice (non-fuzzy sau negraduale).

Capitolul 3

Noi modele de învăţare nesupervizată în extragerea cunoştinţelor din date de natură academică

În acest capitol prezentăm contribuţiile noastre la dezvoltarea modelelor de învăţare nesupervizată pentru analiza datelor de performanţă academică a studenţilor. Secţiunea 3.1 evaluează reţele cu auto-organizare şi modele de extragere a regulilor de asociere relaţională în ceea ce priveşte potenţialul lor de a detecta tipare relevante din seturi de date academice ale studenţilor, mai precis performanţele studenţilor reflectate în notele obţinute pe parcursul unui semestru. Secţiunea 3.2 investighează algoritmul *IRARM* (*extragerea incrementală a regulilor de asociere relaţională)* în contextul datelor educaţionale. Experimentele sunt efectuate pe patru seturi de date şi evidenţiază potenţialul şi relevanţa extragerii incrementale a regulilor de asociere relaţională în descoperirea cunoştinţelor semnificative din datele academice. Abordările prezentate în capitolul curent sunt lucrări de cercetare originale publicate în [9], [8] şi [10].

3.1 Utilizarea modelelor de învăţare nesupervizată pentru analiza performanţei academice a studenţilor

În această secţiune investigăm utilitatea a două tehnici de *învăţare automată nesupervizată* (*reţele cu auto-organizare* şi *extragerea regulilor de asociere relaţională din cunoştinţe*) în analiza performanţei academice a studenţilor, cu scopul mai larg de a dezvolta noi modele de învăţare automată pentru predicţia performanţei studenţilor. Rezultatele experimentale obţinute prin aplicarea modelelor de învăţare nesupervizată menţionate mai sus pe un set de date reale colectate de la Universitatea Babeş-Bolyai subliniază eficacitatea lor în extragerea relaţiilor şi regulilor relevante din datele educaţionale care poate fi utilă pentru a prezice performanţa academică a studenţilor.

Abordarea din această secţiune a fost introdusă în lucrările noastre originale [10] şi [8].

RAR-urile nu au fost utilizate până în prezent pentru analiza seturilor de date academice. În conformitate cu cunoştinţele noastre, un studiu similar nu a fost efectuat în literatura de specialitate.

Această secţiune a examinat două modele de învăţare nesupervizată, *reţele cu auto-organizare* şi *extragerea regulilor de asociere relaţională din cunoştinţe*, în contextul analizei seturilor de date referitoare la performanţa studenţilor. Experimentele efectuate pe un set de date real colectate de la Universitatea Babeş-Bolyai, România, au evidenţiat potenţialul instrumentelor de extragere a cunoştinţelor bazate pe învăţare nesupervizată pentru a detecta şabloane relevante cu privire la performanţa academică a studenţilor. Studiul curent reprezintă punctul de plecare al cercetărilor noastre ulterioare în vederea obţinerii unei mai bune înţelegeri a proceselor de învăţare ale studenţilor şi a dezvoltării unor noi metode de învăţare automată pentru prezicerea performanţelor academice ale studenţilor.

Cercetări viitoare vor fi realizate pentru a extinde evaluarea experimentală la alte seturi de date academice şi pentru a interpreta RAR-urile interesante extrase. Pentru creşterea performanţei procesului de învăţare nesupervizat, vom continua cercetarea metodelor de detectare a anomaliilor şi a instanţelor aberante în date. În plus, vom analiza o fază post-procesare pentru filtrarea setului de RAR-uri extrase pentru eliminarea regulilor care se suprapun în mai multe clase. În plus, un model de clasificare bazat pe RAR va fi avut în vedere pentru a prezice performanţa academică a studenţilor. De asemenea, vom cerceta alte tehnici de învăţare automată bazate pe *clasificare* şi *regresie* pentru predicţia performanţei studenţilor (de exemplu păduri aleatoare de arbori).

3.2 Extragerea incrementală a regulilor de asociere relaţională din seturi de date de natură academică

Extragerea incrementală a regulilor de asociere relaţională ($IRARM$) a fost introdusă anterior ca o metodă eficientă de *extragere online a cunoştinţelor din date* pentru extragerea dinamică a *regulilor de asociere relaţională* (RAR-uri) interesante dintr-un set de date dinamic (si anume un set de date care este extins cu noi instanţe). Studiul realizat în această secţiune a fost introdus în lucrarea originală [9] şi are ca scop să sublinieze eficacitatea RAR şi $IRARM$ ca metode de extragere în contexte de *extragere a cunoştinţelor din date de natură academică*. Experimentele efectuate pe diverse seturi de date academice evidenţiază potenţialul utilizării *regulilor de asociere relaţională* pentru descoperirea cunoştinţelor relevante din date academice.

În această secţiune abordăm problema *extragerii incrementale a regulilor de asociere relaţională* ($IRARM$) în contextul EDM. Procesul de extragere incrementală a RAR-urilor este adecvat în special pentru scenariile DM *online*, unde setul de date care urmează să fie extras este dinamic şi astfel este extins continuu cu fluxuri de date sosite în timp real. În astfel de situaţii, abordarea $IRARM$ urmăreşte adaptarea progresivă a RAR-urilor interesante identificate într-un set de date, atunci când este extins cu noi instanţe. Întrucât procesele de învăţare din mediile educaţionale sunt prin natură procese online, ideea investigării perspectivei $IRARM$ în EDM apare în mod natural. Literatura EDM dezvăluie şi ea că DM este foarte utilă în domeniul educaţional, în special atunci când se explorează mediul de învăţare online [29].

Contribuţia secţiunii este rezumată după cum urmează. În primul rând, subliniem relevanţa extragerii RAR-urilor în domeniul *extragerii cunoştinţelor din date de natură academică* (EDM) cu scopul de a descoperi tipare semnificative în cadrul seturilor de date educaţionale. În al doilea rând, extindem evaluarea experimentală a *extragerii incrementale a regulilor de asociere relaţională* ($IRARM$) [25] pentru mai multe studii de caz EDM. Eficacitatea

IRARM este subliniată prin reducerea timpului de extragere obţinut atunci când se utilizează *IRARM* comparativ cu extragerea RAR-urilor de la zero când setul de date este extins cu noi instanţe. Studiul realizat în această lucrare este inedit în literatura de specialitate EDM, întrucât nici abordarea prin extragerea incrementală a RAR-urilor şi nici cea prin extragerea clasică a acestora nu au fost aplicate pe seturi de date academice până în prezent.

În această secţiune am investigat aplicarea extragerii clasice şi incrementale a RAR-urilor pentru descoperirea cunoştinţelor în seturi de date din medii educaţionale, cu scopul de a descoperi şabloane semnificative din seturi de date educaţionale. Relevanţa descoperirii RAR-urilor în seturile de date academice a fost accentuată în contextul procesului de învăţare a studenţilor, oferind informaţii suplimentare asupra fenomenelor legate de educaţie. În plus, eficacitatea extragerii *incrementale* a RAR-urilor în scenariile EDM online a fost evidenţiată prin mai multe studii de caz.

Cercetări viitoare vor fi efectuate pentru a extinde evaluarea experimentală a *IRARM* la alte sarcini EDM, pentru a-i testa în continuare performanţa. Va fi cercetată extragerea *incrementală adaptativă* a RAR-urilor pentru seturi de date academice, în cazul în care la setul de date sunt adăugate atât instanţe noi, cât şi funcţii noi. În plus, intenţionăm să aplicăm RAR, RAR gradual [17] şi algoritmul de extragere *IRARM* în scenarii EDM de învăţare supervizată, cum ar fi prezicerea performanţei academice a studenţilor.

Capitolul 4

Modele de învăţare supervizată pentru predicţia performanţei studenţilor

În acest capitol prezentăm contribuţiile noastre în dezvoltarea modelelor de învăţare supervizată pentru predicţia performanţei studenţilor.

Efectuăm în [13] o analiză a metodelor de învăţare supervizată pentru predicţia performanţei studenţilor în medii academice. Experimentele sunt efectuate pe trei seturi de date academice reale şi investighează eficacitatea a două modele de regresie *păduri aleatoare de arbori* (RF) şi *reţele neuronale artificiale* (ANN) în predicţia performanţei academice a studenţilor. În lucrarea de cercetare [15] introducem un nou model de clasificare numit $SPRAR$ (*predicţia performanţei studenţilor folosind reguli de asociere relaţională*). $SPRAR$ este utilizat pentru a prezice rezultatele finale obţinute de studenţi la un anumit curs academic, pe baza notelor lor primite în timpul semestrului. Evaluarea este realizată pe trei seturi de date academice reale colectate de la Universitatea Babeş-Bolyai din România. Se compară clasificatorul $SPRAR$ din punct de vedere al performanţei cu clasificatorii supervizaţi din lucrări existente. Experimentele efectuate evidenţiază că $SPRAR$ depăşeşte predictorii performanţei studenţilor propuşi anterior într-o mare majoritate a cazurilor. Lucrarea originală fost publicată în [15].

Abordările prezentate în capitolul curent sunt lucrări de cercetare originale publicate în [13], [15] şi [11].

4.1 O analiză a metodelor de învăţare supervizată pentru predicţia performanţei studenţilor în medii academice

Această secţiune analizează performanţa modelelor de clasificare şi regresie supervizate pentru predicţia performanţei academice a studenţilor. Investigăm eficienţa diferitelor modele de învăţare automată supervizată şi ne propunem să prezicem nota finală a studenţilor la o anumită disciplină academică pe baza performanţei lor din timpul semestrului. Prezentarea din această secţiune se bazează pe lucrarea originală [13].

În această secţiune realizăm un studiu privind aplicarea a două modele de regresie, *păduri aleatoare de arbori* (RF) şi *reţele neuronale artificiale* (ANN), pentru predicţia performanţei academice a studenţilor. Pentru fiecare regresor, sunt prezentate câte două modele de calcul (unul de regresie şi unul de clasificare) pentru a prezice nota studentului la examenul final pe baza notelor sale primite în timpul unui semestru academic. Rezultatele obţinute pe trei

seturi de date reale colectate de la Universitatea Babeş-Bolyai, România dezvăluie că regresorii supervizaţi sunt utili pentru identificarea mijloacelor de creştere a calităţii proceselor educaţionale. În conformitate cu cunoştinţele noastre, un studiu similar nu a fost realizat până acum în literatura EDM. Clasificatorii RF au fost aplicaţi în literatură, dar în alte sarcini decât cea avută în vedere în această lucrare.

Pentru a rezuma, scopul studiului realizat în această secţiune este de a răspunde la următoarele întrebări de cercetare:

RQ1 Care este potenţialul modelelor de regresie supervizată pentru a prezice nota de la examenul final al studenţilor pe baza notelor primite în timpul semestrului?

RQ2 În ce măsură selectarea caracteristicilor îmbunătăţeşte acurateţea prezicerii performanţei studenţilor?

RQ3 Cum se compară modelele de învăţare supervizată utilizate în această lucrare cu alte lucrări conexe din literatura de specialitate în termeni de predicţie a performanţei?

Studiul din această secţiune a fost realizat cu scopul de a evidenţia eficacitatea modelelor de regresie supervizată în prezicerea performanţei academice a studenţilor.

Experimentele efectuate pe trei seturi de date conţinând date academice reale colectate de la o universitate din România au subliniat că, în general, RF sunt cel mai bun model de regresie pentru a prezice nota finală a studenţilor pe baza notelor primite în timpul unui semestru academic. Pentru experimentul efectuat am observat că regresorul SVM furnizează rezultate uşor mai bune decât RF. Studiul a relevat, de asemenea, dificultatea sarcinii de predicţie a performanţei studenţilor şi importanţa creşterii numărului de caracteristici utilizate în procesul de învăţare.

Cercetările viitoare vor urmări adăugarea mai multor caracteristici la sarcinile noastre de învăţare, precum şi aplicarea unor tehnici de preprocesare pentru a detecta instanţele aberante din date. De asemenea, ne propunem să investigăm utilizarea altor modele de învăţare pentru predicţia performanţelor finale ale studenţilor, cum ar fi un clasificator bazat pe *extragerea regulilor de asociere relaţională*.

4.2 *SPRAR*: Un model de clasificare bazat pe extragerea regulilor de asociere relaţională aplicat pentru predicţia performanţei academice

Această secţiune abordează problema predicţiei performanţei academice a studenţilor, o problemă intens investigată în literatura privind extragerea cunoştinţelor din date de natură academică (EDM). Pentru o mai bună înţelegere a fenomenelor legate de educaţie, există un interes continuu în aplicarea metodelor de învăţare *supervizată* şi *nesupervizată* pentru obţinerea unor informaţii suplimentare despre procesul de învăţare al studenţilor. Problema prezicerii faptului că un student va promova sau nu la o anumită disciplină academică bazată pe notele studenţilor din timpul semestrului este una dificilă, şi depinde foarte mult de diverse condiţii, cum ar fi disciplina (cursul), numărul de examene pe semestru, profesorii şi exigenţele lor. Problema are o relevanţă practică majoră în mediile educaţionale, deoarece ar putea oferi feedback relevant studenţilor care ar putea să nu promoveze la un anumit curs academic. Având un astfel de sfat în timpul semestrului, studenţii vor avea posibilitatea de a preveni un posibil eşec academic. Propunem un nou model de clasificare, *SPRAR* (*predicţia*

performanţei studenţilor folosind reguli de asociere relaţională) pentru a prezice rezultatul final al unui student la o anumită disciplină academică folosind *reguli de asociere relaţională* (RAR). Abordarea $SPRAR$ a fost introdusă în lucrarea originală [15]. Performanţa clasificatorului $SPRAR$ în studiile de caz considerate este comparată cu cea din lucrările existente, fiind superioară predictorilor de performanţă a studenţilor propuşi anterior. Trebuie să notăm generalitatea clasificatorului $SPRAR$, care nu este specific sarcinii de predicţie a performanţei studenţilor. $SPRAR$ poate fi adaptat cu uşurinţă la alte probleme de clasificare.

Modelul de clasificare introdus este unul binar (adică există doar două clase care pot fi prezise: *va promova* sau *nu va promova*), dar modelul propus este unul general, şi poate fi extins pentru o problema de clasificare cu clase multiple (de exemplu, predicţia notei finale a studentului). Dintr-o perspectivă de învăţare supervizată, problema predicţiei finalizării cu succes a unui curs în timpul unui semestru academic este o problemă dificilă, în special datorită naturii dezechilibrate a datelor de antrenament (de exemplu numărul de studenţi care au promovat examenul este în general mult mai mare decât numărul de studenţi care *nu au promovat* examenul). Experimente efectuate pe trei seturi de date reale de la Universitatea Babeş-Bolyai, România evidenţiază eficacitatea clasificării folosind $SPRAR$. Literatura de specialitate cu privire la predicţia performanţei studenţilor relevă faptul că utilizarea RAR-urilor în prezicerea performanţei academice a studenţilor reprezintă o abordare inedită.

În concluzie, scopul cercetării efectuate este propunerea unui model de clasificare $SPRAR$ bazat pe extragerea RAR pentru a prezice performanţa academică a studenţilor, ca dovadă a conceptului. În acest scop, dovada conceptului ia în considerare doar trei seturi de date de dimensiuni medii pentru a evidenţia faptul că $SPRAR$ este potrivit pentru problema abordată. Dacă acest lucru este valabil, studiul aplicării $SPRAR$ pentru predicţia performanţei studenţilor poate fi extins în continuare la o scară mai mare.

Cercetările viitoare vor urmări extinderea evaluării experimentale a $SPRAR$ la alte studii de caz. În plus, ne propunem să generalizăm clasificatorul binar $SPRAR$ la un clasificator cu mai multe clase, astfel încât să putem prezice nota finală a studenţilor la un anumit curs academic. În ceea ce priveşte procesul de descoperire a regulilor de asociere relaţională, intenţionăm să extindem modelul nostru, luând în considerare reguli de *asociere relaţională graduală* [17]. Vor fi studiate în continuare măsuri alternative pentru definirea probabilităţilor folosite în procesul de clasificare a unei instanţe, precum şi ideea folosirii unui ansamblu de clasificatori $SPRAR$ pentru îmbunătăţirea performanţei predictive. $SPRAR$ poate fi extins şi pentru gestionarea relaţiilor n-are dintre valorile atributelor.

4.3 Un studiu privind predicţia performanţei studenţilor folosind $SPRAR$

Această secţiune analizează modelul de clasificare SPRAR (predicţia performanţei studenţilor folosind reguli de asociere relaţională) introdus în [15], Secţiunea 4.2, pentru a prezice rezultatele academice ale studenţilor utilizând reguli de asociere relaţională. Trei noi scoruri de clasificare sunt introduse în acest capitol şi sunt utilizate în etapa de clasificare a SPRAR. Performanţa lor este analizată folosind trei seturi de date academice reale de la Universitatea Babeş-Bolyai, România şi comparate cu rezultatele similare existente în literatura de specialitate.

Materialul prezentat în această secţiune este o lucrare originală publicată în [11]. În această lucrare am extins modelul de clasificare $SPRAR$ prin introducerea a trei scoruri

alternative care pot fi utilizate în etapa de clasificare a $SPRAR$ pentru a decide dacă un student *va promova* sau *nu va promova* la un anumit curs. Au fost utilizate trei studii de caz reale pentru evaluarea performanţei $SPRAR$ având în vedere scorurile de clasificare recent propuse. Cu scopul de a determina empiric care este scorul de clasificare cel mai potrivit, scorurile propuse sunt analizate comparativ şi evaluate în raport cu varianta iniţială introdusă în [15]. Studiul efectuat în această secţiune este nou în literatura EDM.

Această secţiune a extins modelul de clasificare SPRAR (predicţia performanţei studenţilor folosind reguli de asociere relaţională) [15] pentru a prezice realizările studenţilor la un anumit curs bazat pe extragerea *regulilor de asociere relaţională* (RAR) [3]. Apartenenţa unei instanţe de interogare la o clasă de trecere sau eşec este determinată folosind seturi de date academice obţinute de la Universitatea Babeş-Bolyai, România. Trei noi scoruri de clasificare sunt introduse şi utilizate în etapa de clasificare a SPRAR pentru a decide dacă un student *va promova* sau *nu va promova*. Performanţa SPRAR este evaluată folosind mai multe măsuri de evaluare: *sensibilitate*, *specificitate* şi *aria de sub curba ROC*. Cel mai bun scor de predicţie este identificat şi comparat cu rezultatele existente din literatura actuală.

Prin mai multe experimente efectuate pe date academice reale, am obţinut o dovadă empirică conform căreia regulile de asociere relaţională sunt relevante pentru a face diferenţa între performanţa academică a studenţilor. De asemenea, am observat că metodologia de clasificare (identificarea celui mai bun scor pentru a discrimina între performanţa studenţilor) depinde de seturile de date şi prin urmare am putea investiga în continuare posibilitatea de a o învăţa automat, folosind metode supervizate. Ca viitoare direcţii de cercetare ne propunem, de asemenea, să extindem experimentele noastre pe alte seturi de date, pentru a adăuga atribute şi relaţii suplimentare în procesul de extragere RAR pentru îmbunătăţirea performanţei predictive a SPRAR.

Concluzii

Teza actuală a subliniat principalele contribuţii ale noastre la dezvoltarea şi aplicarea modelelor de extragere a cunoştinţelor din date pentru rezolvarea unor probleme legate de domeniul educaţional. Extragerea cunoştinţelor din date de natură academică (EDM) este un domeniu de cercetare în curs de dezvoltare care se concentrează pe dezvoltarea unor modele pentru identificarea şi extragerea informaţiilor relevante din datele academice. Cercetarea în EDM îşi propune să descopere şabloane în procesul de învăţare a studenţilor şi să îmbunătăţească realizările lor academice. Succesul şi performanţa academică a studenţilor pot fi prezise prin tehnici de extragere a cunoştinţelor din date [20] folosind învăţarea *supervizată* sau *nesupervizată*.

Regulile de asociere relaţională (RAR-uri) [5, 3] pot exprima diferite tipuri de relaţii neordinale dintre atributele datelor. Atât extragerea regulilor de asociere ordinale, cât şi a celor relaţionale au fost aplicate cu succes pentru a soluţiona probleme care variază de la sarcinile de extragere a cunoştinţelor din date (de exemplu, detectarea defectelor de proiectare software [18] sau curăţarea datelor [5]) la clasificarea supervizată (de exemplu, predicţia defectelor software [26]).

Am introdus o nouă metodă de extragere incrementală a regulilor de asociere relaţională, numită $IRARM$, care extrage eficient toate RAR-urile interesante dintr-un set de date dinamic. Deoarece procesul de extragere a RAR-urilor este costisitor din punct de vedere computaţional, ne-am concentrat de asemenea pe dezvoltarea unei versiuni concurente a algoritmului de extragere RAR şi anume $CRAR$, cu obiectivul principal de reducere a timpul de calcul al procesului de extragere.

Modelele de învăţare nesupervizată şi supervizată au fost, de asemenea, studiate în teză în contextul EDM şi anume: extragerea incrementală RAR şi reţelele cu auto-organizare pentru analiza datelor de performanţă ale studenţilor. Experimentele efectuate pe mai multe seturi de date educaţionale reale au evidenţiat potenţialul metodelor de învăţare nesupervizată în descoperirea unor tipare semnificative în datele academice ale studenţilor. Ca dovadă a conceptului, am introdus un nou model de clasificare $SPRAR$ [15] bazat pe descoperirea *regulilor de asociere relaţională* pentru a prezice finalizarea cu succes a unui curs academic, în funcţie de notele primite de studenţi în timpul semestrului academic. Au fost introduse şi utilizate trei scoruri de clasificare în etapa de clasificare a SPRAR pentru a decide dacă un student *va promova* sau *nu va promova* examenul [11].

În ceea ce priveşte contribuţiile noastre conceptuale în domeniul DM, ne propunem să investigăm condiţiile exacte în care este mai eficient să se aplice metoda incrementală $IRARM$ decât să se ruleze algoritmul offline $DRAR$ de la zero. În plus, intenţionăm să investigăm o abordare incrementală adaptativă a extragerii *regulilor de asociere relaţională*, în care sunt adăugate atât atribute noi, cât şi obiecte noi la setul de date extrase. Preconizăm optimizări ulterioare ale metodei extragerii concurente RAR ($CRAR$), cum ar fi: creşterea volumului de rulare paralelă a $CRAR$, în special calculul RAR-urilor binare, care ar creşte *eficienţa*

scalării şi *accelerarea paralelă.*

În ceea ce priveşte clasificatorul $SPRAR$ pentru *predicţia performanţei studenţilor*, cercetări viitoare vor fi efectuate pentru a extinde evaluarea experimentală a $SPRAR$ la alte studii de caz. În plus, ne propunem să generalizăm clasificatorul binar $SPRAR$ către un clasificator cu mai multe clase, astfel încât să prezică notele studenţilor la examenul final la un anumit curs academic. În ceea ce priveşte procesul de descoperire a regulilor de asociere relaţională, intenţionăm să extindem modelul nostru având în vedere *reguli de asociere relaţională graduală* [17]. $SPRAR$ poate fi extins şi pentru gestionarea relaţiilor n-are dintre valorile atributelor.

Bibliografie

[1] Alejandro Bogarín, Rebeca Cerezo, and Cristóbal Romero. A survey on educational process mining. *Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery*, 8(1), 2018.

[2] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001.

[3] Gabriela ¸Serban, Alina Câmpan, and Istvan Gergely Czibula. A programming interface for finding relational association rules. *International Journal of Computers, Communications & Control*, I(S.):439–444, June 2006.

[4] Alina Câmpan, Gabriela ¸Serban, and Andrian Marcus. Relational association rules and error detection. *Studia Universitatis Babes-Bolyai Informatica*, LI(1):31–36, 2006.

[5] Alina Câmpan, Gabriela Şerban, and Andrian Marcus. Relational association rules and error detection. *Studia Universitatis Babes-Bolyai Informatica*, LI(1):31–36, 2006.

[6] Alina Campan, Gabriela Şerban, Traian Marius Truta, and Andrian Marcus. An algorithm for the discovery of arbitrary length ordinal association rules. In *DMIN*, pages 107–113, 2006.

[7] H. Y. Chang, J. C. Lin, M. L. Cheng, and S. C. Huang. A novel incremental data mining algorithm based on fp-growth for big data. *2016 International Conference on Networking and Network Applications (NaNA)*, pages 375–378, 2016.

[8] George Ciubotariu and Liana Maria Crivei. Analysing the academic performance of students using unsupervised data mining. *Studia Universitatis Babes-Bolyai Series Informatica*, pages 1–14, 2019.

[9] Liana Maria Crivei. Incremental relational association rule mining of educational data sets. *Studia Universitatis Babes-Bolyai Series Informatica*, 63(2):102–117, 2018.

[10] Liana Maria Crivei. Using unsupervised learning models for analyzing students' academic performance. In *PhD Coloquium at SYNASC'18, Symbolic and Numeric Algorithms for Scientific Computing*, pages 1–4, 2018.

[11] Liana-Maria Crivei, Mihai Andrei, and Czibula Gabriela. A study on applying relational association rule mining based classification for predicting the academic performance of students. In *KSEM 2019 : The 12th International Conference on Knowledge Science, Engineering and Management, LNAI 11775*, pages 287–300, 2019.

[12] Liana Maria Crivei, Gabriela Czibula, George Ciubotariu, and Mariana Dindelegan. Unsupervised learning based mining of academic data sets for students' performance

analysis. In *IEEE 14th International Symposium on Applied Computational Intelligence and Informatics (SACI 2019)*, pages 1–6. IEEE Hungary Section, 2019.

[13] L.M. Crivei, V.-S. Ionescu, and G. Czibula. An analysis of supervised learning methods for predicting students' performance in academic environments. *ICIC Express Letters*, 13(3):181–190, 2019.

[14] G. Czibula, I.G. Czibula, D.-L. Miholca, and L.M. Crivei. A novel concurrent relational association rule mining approach. *Expert Systems with Applications*, 125(6):142–156, 2019.

[15] G. Czibula, A. Mihai, and L.M. Crivei. A novel relational association rule mining classification model applied for academic performance prediction. volume 159, pages 20–29, 2019.

[16] Gabriela Czibula, Maria-Iuliana Bocicor, and Istvan Gergely Czibula. Promoter sequences prediction using relational association rule mining. *Evolutionary Bioinformatics*, 8:181–196, 04 2012.

[17] Gabriela Czibula, Istvan Gergely Czibula, and Diana-Lucia Miholca. Enhancing relational association rules with gradualness. *International Journal of Innovative Computing, Communication and Control*, 13:289–305, 2017.

[18] Gabriela Czibula, Zsuzsanna Marian, and István Gergely Czibula. Detecting software design defects using relational association rule mining. *Knowledge and Information Systems*, 2014.

[19] N. Elfelly, J.-Y. Dieulot, and P. Borne. A neural approach of multimodel representation of complex processes. *International Journal of Computers, Communications & Control*, III(2):149–160, 2008.

[20] Richard A Huebner. A survey of educational data-mining research. *Research in Higher Education Journal*, 19:1–13, 2013.

[21] Andreas Khler, Matthias Ohrnberger, and Frank Scherbaum. Unsupervised feature selection and general pattern discovery using self-organizing maps for gaining insights into the nature of seismic wavefields. *Computers Geosciences*, 35(9):1757 – 1767, 2009.

[22] J. Lampinen and E. Oja. Clustering properties of hierarchical self-organizing maps. *Journal of Mathematical Imaging and Vision*, 2(3):261–272, 1992.

[23] Kristopher R. Linstrom and A. John Boye. A neural network prediction model for a psychiatric application. *International Conference on Computational Intelligence and Multimedia Applications*, 0:36–40, 2005.

[24] Andrian Marcus, Jonathan I. Maletic, and King-Ip Lin. Ordinal association rules for error identification in data sets. In *Proceedings of the tenth international conference on Information and knowledge management*, CIKM '01, pages 589–591, New York, NY, USA, 2001. ACM.

[25] Diana-Lucia Miholca, Gabriela Czibula, and Liana Maria Crivei. A new incremental relational association rules mining approach. In *22nd International Conference on*

Knowledge-Based and Intelligent Information & Engineering Systems, volume 126 of *KES2018*, pages 19–28. Procedia Computer Science, 2018.

[26] Diana-Lucia Miholca, Gabriela Czibula, and István Gergely Czibula. A novel approach for software defect prediction through hybridizing gradual relational association rules with artificial neural networks. *Inf. Sci.*, 441:152–170, 2018.

[27] Diana-Lucia Miholca, Gabriela Czibula, and Istvan Gergely Czibula. A novel approach for software defect prediction through hybridizing gradual relational association rules with artificial neural networks. *Information Sciences*, 441:152 – 170, 2018.

[28] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.

[29] Siti Khadijah Mohamad and Zaidatun Tasir. Educational data mining: A review. *Procedia - Social and Behavioral Sciences*, 97:320 – 324, 2013. The 9th International Conference on Cognitive Science.

[30] Siti Khadijah Mohamad and Zaidatun Tasir. Educational data mining: A review. *Procedia - Social and Behavioral Sciences*, 97(Supplement C):320 – 324, 2013. The 9th International Conference on Cognitive Science.

[31] Mark D. Skowronski and John G. Harris. Automatic speech recognition using a predictive echo state network classifier. *Neural Networks*, 20(3):414–423, April 2007.

[32] Panu Somervuo and Teuvo Kohonen. Self-organizing maps and learning vector quantization for feature sequences. *Neural Processing Letters*, 10:151–159, 1999.

[33] Anping Song, Xuehai Ding, Mingbo Li, Wei Cao, and Ke Pu. A novel binary bat algorithm for association rules mining. *ICIC Express Letters*, 9(9):2387–2394, September 2015.

[34] Ruihu Wang. Adaboost for feature selection, classification and its relation with svm, a review. *Physics Procedia*, 25:800 – 807, 2012. International Conference on Solid State Devices and Materials Science, April 1-2, 2012, Macao.

[35] Tiantian Yang, Ata Akabri Asanjan, Mohammad Faridzad, Negin Hayatbini, Xiaogang Gao, and Soroosh Sorooshian. An enhanced artificial neural network with a shuffled complex evolutionary global optimization with principal component analysis. *Information Sciences*, 418-419(Supplement C):302 – 316, 2017.