Universitatea Babeș-Bolyai Cluj-Napoca
Facultatea de Matematică și Informatică

MODELE DE CLASIFICARE PENTRU PROBLEME DE NATURĂ MEDICALĂ ȘI PSIHOLOGICĂ

Rezumat teza de doctorat – în limba română

Autor:
Adriana Mihaela COROIU

Coordonator științific:
Prof. Dr. Horia F. POP

Cluj-Napoca
2018

1 INTRODUCERE

Această lucrare este realizată într-un moment al tehnologizării celor mai multe dintre activitățile standard din viața de zi cu zi. Inteligența Artificială (*IA*) și tot ceea ce oferă acestă parte a informaticii a evoluat într-un ritm rapid și noi avem constant parte de exemple care să certifice acest lucru. In plus, inovația din acest domeniu devine din ce in ce mai importantă și de un real ajutor pentru persoanele care aleg sau care pot să beneficieze de ceea ce oferă acest domeniu de cercetare atât de productiv.

Utilizarea avantajelor oferite de către *IA* în domeniul științelor medicale devine din ce in ce mai necesară și oferă rezultate tot mai bune, imposibil de atins fără utilizarea tehnicilor și instrumentelor oferite de către domeniul informaticii. Astfel, ultimele noutăți rezultate din combinarea celor două domenii sunt:

- *crearea unor micro-computere în interiorul celulelor vii* care studiază sistemul de viață al organismului și chiar să il reprogrameze în anumite contexte;
- *crearea unui "robot orb"* (fără cameră sau senzori vizuali) de către cercetătorii de la MIT care reușeste să detecteze drumul optim, depășind toate obstacolele întâlnite;
- *tratarea fenomenului de orbire al corneei* prin realizarea unei membrene similare țesutului natural al ochiului folosind imprimantele 3D.

Cele menționate mai sus arată încă o data importanța utilizării *IA* in domeniul științelor medicale și susțin relevanța acestei lucrări.

Lucrarea tratează aplicabilitatea tehnicilor din subdomeniul Inteligenței Artificiale, numit *Machine Learning* (*ML*) în două domenii cheie pentru omenire: medicină, respectiv psihologie.

Obiectivul realizării acestei lucrări de doctorat este de a identifica și prezenta modul în care *ML* (și implicit *IA*) are implicații în stabilirea unui diagnostic, mai precis în îmbunătățirea deciziei pentru un anumit diagnostic dar și de a descoperi șabloane în seturile de date analizate și dependențele dintre acestea cât și metode care determină rezultate cu o bună precizie pentru noi instanțe ale seturilor de date.

Fiecare capitol din lucrare se concentrează pe aspecte diferite atât din perspectiva tehnicilor folosite cât și din perspectiva problemelor de natura medicală sau psihologie abordate. Sunt comparate metode diferite de analiză și de predicție cu scopul de a vedea care dintre acestea ne oferă rezultate mai bune pentru problema particulară abordată. Rezultatele obținute în urma analizelor efectuate sunt promițătoare și din punctul nostru de vedere pot să ajute un specialist în luarea unei decizii. Considerăm util acest lucru, întrucât rezultatele analizelor se bazează pe seturi de date cu mai multi pacienți, iar capacitatea de sinteză a instrumentelor tehnice este mult mai ridicată atunci când există multe informații și, de asemenea, dispare subiectivitatea.

Ca și viitoare direcții de cercetare, aceste domenii permit dezvoltarea de sisteme noi care să faciliteze mare parte dintre procesele esențiale necesare vieții și societății. Pentru a ne da seama mai bine de importanța *IA* în viața noastră, aducem în discuție un articol publicat recent de către ziarul Harvard Business, unde unul dintre cei mai influenți cercetători din domeniul Inteligenței Artificiale, Andrew Ng afirmă că dacă momentan există o activitate efectuată de către o persoană într-un timp aproximativ de o secundă, cel mai probabil, acea activitate va fi automatizată de *IA* fie acum, fie in viitorul apropiat.

2 STRUCTURA LUCRĂRII

Lucrare este structurată pe trei capitole, fiecare capitol este descris insistând pe prezentarea elementelor definitorii și pe abordările propuse, precum și pe rezultatele obținute.

Capitolul **Preliminarii Teoretice** prezintă care sunt abordările istorice ale învățării automate în diferite domenii cu accent pe domeniile medical și psihologic, dar nu numai. De asemenea, este prezentat modul de lucru interdisciplinar raportându-ne la literatura de specialitate existentă până în acest moment.

Capitolul **Clasificarea datelor numerice și multicategoriale. Aplicații în medicină** este format din trei părți altfel:

- prima parte tratează o problemă de clasificare nesupervizată având la bază cele mai relevante metode de *linkage* și modul în care acestea produc rezultate (adică gruparea datelor în clustere) și evaluarea acestora folosind metode interne și externe de evaluare;
- a doua parte are în vedere determinarea celor mai relevante atribute indicate să fie selectate astfel încât rezultatele determinate în urma aplicării metodelor de clasificare să aibă o precizie cât mai bună;
- ultima parte a acestui capitol are ca și scop tot îmbunătățirea metodelor de clasificare prin selectarea valorilor optime ale parametrilor metodelor de clasificare.

Capitolul **Clasificarea datelor binare. Aplicații în psihologie** este împărțit în două părți, din perspectiva seturilor de date analizate:

- în primul rând, chestionarul pentru analiza stilului de comunicare (SCq). Răspunsurile la acest chestionar au dus la formarea unui set de date cu 220 de instanțe (adică numărul de pacienți) și 63 de atribute (60 dintre ele fiind răspunsuri la întrebările chestionarului propriu-zis iar alte 3 atribute cuantifică alte informații despre participanții la chestionar: vârsta persoanei, sexul persoanei și stresul auto-perceput). Pentru acest set de date am analizat probabilitatea de predictibilitate cu succes a clasei de determinat (variabila țintă) după aplicarea metodelor de clasificare.
- în al doilea rând, chestionarul de coping cognitiv-emoţional adaptat (tCERQ). Răspunsurile pacienților la acest chestionar au condus la crearea unui set de date de 240 de instanțe și 36 de atribute, de tip binar asupra căruia au fost aplicate metode de detecție a datelor cu comportament anormal.

  In plus am dorit să verificăm dacă există anumite corelații între rezultatele obținute de către noi, în consecință pentru ambele seturi de date analizate în acest capitol am avut în vedere și aplicarea tehnicilor pentru identificarea corelațiilor existente în seturile de date.

Capitolul **Identificarea opiniilor prin analiza de text** tratează aspecte legate de subdomeniul *Opinion Mining* (OM) cu aplicabilitate asupra unui set

de date colectat de către autor care constă în opinii ale studenților cu privire la modul de desfășurare al activității de seminar. Scopul acestei părți a fost de a determina dacă studenții au o opinie pozitivă sau negativă despre modul de desfășurare al seminarului fără să li se ceară în mod deschis acest lucru. Setul de date analizat este un set de date cu 180 de instanțe (reprezentând numărul de studenți care și-au exprimat opinia) și cu un singur atribut de tip text reprezentând răspunsul la întrebarea "Cum a fost seminarul de-a lungul semestrului, din perspectiva voastră?"

In final, capitolul **Concluzii și direcții viitoare de cercetare** trasează ideile și rezultatele care au dus la realizarea acestei lucrări cât și propunerea altor direcții de lucru, fie prin îmbunătățirea aspectelor deja prezentate fie prin propunerea unor idei și metode de lucru inovative.

3 CONTRIBUȚII ORIGINALE

Din perspectiva relevanței lucrării pentru comunitatea științifică, rezultatele obținute de către noi sunt utile și au aplicabilitate în subdomeniile abordate.

Vom sumariza aceste rezultate, pe capitole și vom evidenția utilitatea experimentelor.

- Rezultale obținute în capitolului **Clasificarea datelor numerice și multicategoriale. Aplicații în medicină** sunt:

  1. O analiză experimentală asupra metodelor de linkage în metode de clustering aglomerativ cu scopul de a arăta care dintre metode determină rezultate optime. Evaluarea metodelor de clustering din analiza efectuată de către noi s-a realizat prin măsuri de validare internă și externă. Utilitatea practică a acestei analize este reprezentată de extinderea aplicării ei (cu metodele de linkage și metricile de evaluare descoperite de către noi ca fiind optime) pentru seturi noi de date, cu aceleași tipuri de atribute.
  2. O abordare particulară bazată pe meta-euristica Algoritmilor Genetici pentru determinarea unui subset relevant de atribute, din întregul set existent. Metoda propusă, bazata pe Algoritmii Genetici a fost comparată cu alte metode existente, diferențele fiind evaluate prin aplicarea metodelor de clasificare pentru subseturile obținute prin fiecare metodă cât și pe întregul set. Validarea metodelor de clasificare s-a efectuat prin metrici specifice de cuantificare a peformanței acestor metode. Utilitatea acestor rezultate este reprezentată de existența unei noi abordări pentru determinarea unui subset relevant de atribute, cât și de analiza comparativă existentă.
  3. O metodă pentru selectarea valorilor optime ale parametrilor metodelor de clasificare bazată pe Algoritmi Genetici. Pentru a valida eficiența metodei noastre am realizat experimentele folosind și alte două metode pentru a determina valorile parametrilor metodelor de clasificare. Rezultatele au evidențiat utilitatea metodei propuse prin îmbunătățirea valorii metricii de acuratețe ca măsura a performanței metodelor de clasificare.

- Rezultale obținute în capitolului **Clasificarea datelor binare. Aplicații în psihologie** sunt:

  1. colectarea și crearea unui set de date mixt format pe baza răspunsurilor la chestionarul pentru analiza stilului de comunicare (220 de instanțe și 63 de atribute mixte);
  2. determinarea unei metode optime de clasificare pentru setul nostru particular de date și determinarea corelațiilor existente între stilurile de comunicare ale participanților și nivelul de stres auto-perceput de către aceștia.

3. colectarea și crearea unui set de date format pe baza răspunsurilor la chestionarul de coping cognitiv-emoţional adaptat (240 de instanțe și 36 de atribute binare);
4. determinarea instanțelor cu comportament anormal din acest set de date și evaluarea performanței metodelor de clasificare după eliminarea acestor date;
5. investigarea eficienței metodelor de detecție al datelor cu comportament anormal: Eliptic Envelope, Isolation Forest și Local Outlier Factor, cât și realizarea unei analize comparative al instanțelor anormale detectate de către fiecare metodă;
6. verificarea existenței corelațiilor între strategiile de coping existente în setul de date.

Utilitatea și relevanța acestor rezultate vine din perspectiva originalității seturilor de date colectate, a îmbinării metodelor aplicate cât și din coroborarea acestor rezultate cu aspectele teoretice existente în literatura de specialitate din psihologie.

- Rezultale obținute în capitolului **Identificarea opiniilor prin analiza de text** sunt:

1. colectarea unui set de date cu 180 de instanțe și un atribut de tip text corespunzător răspunsurilor studenților participanți la seminarul de Programare Logică și Funcțională. Aceste răspunsuri sintetizează opinii ale studenților cu privire la modul de desfășurare al seminarului;
2. realizarea unei clasificări a acestor opinii într-un mod obiectiv *(Opinion Mining)* și posibilitatea de a efectua predicții pentru noi instanțe;
3. propunerea și folosirea unei meta-metode de clasificare bazată pe cele mai bune rezultate ale metodelor de clasficare (fiecare metoda de clasificare primeste un grad de încredere pe baza rezultatelor obținute anterior) deja aplicate asupra setul de date.

   In momentul redactării acestei lucrări nu au fost găsite abordări similare în literatura de specialitate, astfel considerăm relevantă abordarea noastră, atât din perspectiva originalității setului de date cât și din perspectiva abordării unui text din perspectiva *Opinion Mining* în contrast cu *Sentiment Analysis*.
4. realizarea unei grupări a cuvintelor care compun conținutul textelelor instanțelor, pe baza tehnicii de *topic modeling*.

Considerăm că o lucrare de acest fel era și este necesară, mai ales în contexul utilizării constante a mecanismelor oferite de domeniul tehnic celorlalte domenii de cercetare sau studiu, fie științific, fie clinic. Cele două domenii abordate în cadrul acestei lucrări de doctorat sunt două domenii reprezentative pentru fiecare dintre noi întrucât sănătatea fizică și mintală

sunt vitale. În acest context, orice mecanisme care să le îmbunătățească sunt binevenite.

4 CUVINTE CHEIE

- **clasificare**
- **date mixte**
- **optimizare parametrii**
- **algoritmi genetici**
- **detecție outliers**
- **opinion mining**

5 CUPRINSUL LUCRĂRII DE DOCTORAT

Cuprins

ABSTRACT i

MULTUMIRI iii

LISTA FIGURILOR viii

LISTA TABELELOR x

LISTA PUBLICATIILOR xi

1 INTRODUCERE **1**

1.1 Motivația alegerii temei 1

1.2 Contribuții de natură informatică în medicină și psihologie 3

1.3 Structura lucrării 4

1.4 Contribuții originale 5

2 PRELIMINARII TEORETICE **10**

2.1 Scurt istoric 10

2.2 Definire și domenii conexe 12

2.3 Cum se ajunge la un model de invățare - aspecte teoretice 16

2.3.1 Preprocesarea seturilor de date 16

2.3.2 Selecția atributelor relevante 20

2.3.3 Selectarea unei metode de învățare și evaluarea unui model 21

2.3.4 Evaluarea modelelor și predicția pentru o nouă instanță 22

2.4 Instrumente soft utilizate 24

3 Clasificarea datelor numerice și multicategoriale. Aplicații în medicină **26**

3.1 Gruparea datelor mixte 26

3.1.1 Descrierea seturilor de date 27

3.1.2 Descrierea metodelor de linkage în clustering 27

3.1.3 Validarea metodelor de clustering . . . 29
3.1.4 Interpretarea rezultatelor . . . 31
3.1.5 Studii comparative pentru metodele de clustering . . . 33
3.2 Imbunătățirea performanței prin detectarea atributelor relevante . . . 33
3.2.1 Extragerea vs. selecția atributelor . . . 33
3.2.2 Descrierea seturilor de date . . . 38
3.2.3 Rezultate și discuții . . . 39
3.2.4 Studii comparative pentru selecția celor mai relevante atribute . . . 41
3.3 Determinarea valorilor parametrilor metodelor de clasificare . . . 41
3.3.1 Descrierea seturilor de date . . . 41
3.3.2 Tehnici de căutare și metode de clasificare . . . 41
3.3.3 Discuții și rezultate . . . 43
3.3.4 Studii comparative pentru hyperparametrizare . . . 44
3.4 Concluzii pentru abordările capitolului . . . 45

4 Clasificarea datelor binare. Aplicații în psihologie 48
4.1 Analiza stilului de comunicare (SCq) . . . 49
4.1.1 Descrierea experimentelor . . . 49
4.1.2 Setul de date . . . 50
4.1.3 Metode de clasificare . . . 52
4.1.4 Evaluarea modelelor . . . 53
4.1.5 Rezultate si discuții . . . 54
4.2 Strategii de coping-cognitiv emoțional - tCERQ . . . 57
4.2.1 Metode de detecție a datelor cu comportament anormal . . . 58
4.2.2 Descrierea experimentului . . . 59
4.2.3 Descrierea setului de date . . . 60
4.2.4 Rezultate și discuții . . . 60
4.3 Concluzii ale capitolului . . . 62

5 Identificarea opiniilor prin analiza de text 65
5.1 Opinion mining versus Sentiment analysis . . . 65
5.2 Noțiuni utilizate . . . 66
5.3 Descrierea setului de date . . . 70

5.4 Clasificarea textului și predicția pentru o nouă instanță 71
5.5 Rezultate și discuții 72
5.6 Imbunătățirea metodelor de clasificare 73
5.7 Studii comparative existente 74
5.8 *Topic modeling* și document clustering 75
5.9 Concluzii ale capitolului 77

6 Concluzii și direcții viitoare de cercetare **79**
6.1 Concluzii 79
6.2 Direcții viitoare de cercetare 81
Referințe 83