Contribuții la îmbunătățirea tehnologiei de clasificare a datelor

- REZUMAT TEZĂ DE DOCTORAT -

Autor:
Asist. Ing. Raul ROBU

Coordonator:
Prof. Dr. Ing. Vasile Stoicu-Tivadar

Clasificarea este una dintre cele mai populare tehnici de data mining. Algoritmii de clasificare au un domeniu de aplicabilitate foarte larg. Ei pot fi utilizați cu succes pentru clasificarea amprentelor, bolilor de inimă, a riscului de reapariție a unor evenimente specifice cancerului de sân, emoțiilor de pe fața umană, imaginilor de la sateliți, etc. Prin intermediul algoritmilor de clasificare o universitate poate prezice cu o precizie între 75 şi 80 % care studenți vor absolvi şi care nu. Universitatea poate folosi această informație pentru a-şi concentra atenția asupra acelor studenți care sunt în pericol de a nu absolvi.

Teza tratează aspectele specifice analizei datelor prin tehnici de clasificare, parcurgând prin structura propusă întreg procesul: preprocesare, construirea modelului, testarea, realizarea de predicții şi ceea ce este cel mai important, toate sugestiile noi de algoritmi sau tehnici sunt validate pe parcursul lucrării sau în capitolul aplicativ 6, prin studii de caz reale şi comparații temeinice privind performanțele.

În capitolul introductiv a fost realizată o sinteză amplă a celor mai importante aspecte legate de domeniul Data Mining. Au fost prezentate definiții şi terminologia folosită în domeniu şi etapele procesului de data mining care au fost redate printr-o schemă concepută de autor. Au fost prezentate 4 domenii în care tehnicile de data mining se aplică cu succes: medicină, educație, combaterea terorismului şi e-commerce cu exemple concrete de realizări în fiecare din aceste domenii. A fost prezentată o scurtă sinteză a raportului Oficiul de Contabilitate Generală al Statelor Unite (GAO - The General Accounting Office) cu privire la utilizarea sau intenția de utilizare a tehnologiei data mining în departamentele şi agențiile federale din SUA. A fost detaliat conceptul de clasificare a datelor şi a fost descris principiul de funcționare a 3 algoritmi clasici utilizați în clasificarea datelor: Naive Bayes- construieşte un model clasificator bazat pe calculul unor probabilități condiționate, k-Nearest Neighbor – clasifică instanțele pe baza clasei majoritare pe care o au cele mai apropiate k instanțe de instanța curentă, C4.5 - construieşte un clasificator ce poate fi reprezentat sub forma unui arbore de decizie în care nodurile reprezintă teste după valorile atributelor iar frunzele reprezintă clasele. Tot în introducere sunt prezentate aspecte generale legate de instrumentele software care pot fi utilizate în studii de data mining. Sunt prezentate aspecte legate de utilizarea instrumentelor Weka şi R şi este realizată şi o comparație între acestea. Instrumentul care se situează pe primul loc în preferințele autorului şi asupra căruia a fost concentrată atenția pe parcursul tezei este Weka. Au fost prezentate detalii privind seturile de date care au fost utilizate în capitolele 3, 4 şi 5 pentru a valida algoritmii şi îmbunătățirile

tehnologice. Seturile de date conţin date medicale despre boli: precum cancerul de sân, boli de inimă, diabet, boli dermatologice, date despre animale, despre planete (flori iris şi ciuperci), despre evaluările unor maşini, despre evaluările aplicaţiilor la o grădiniţă, despre mutările de la jocul X şi zero. Datele sunt reale şi sunt preluate de pe UCI Machine Learning Repository.

Capitolul 2 tratează prima etapă a oricărui studiu de data mining: preprocesarea datelor. Datele iniţiale pot conţine valori lipsă, date introduse greşit, date expirate, date duplicat; în acest sens se impune o primă prelucrare (preprocesare) a acestora prin estimarea sau eliminarea valorilor lipsă, prin corectarea sau eliminarea datelor introduse greşit, eliminarea datelor expirate etc. Datele curăţate este posibil să trebuiască preprocesate în continuare pentru a le aduce în formatul cel mai potrivit pentru algoritmii de data mining care urmează să fie utilizaţi. Pentru aceasta se pot utiliza tehnici precum agregarea (combinarea mai multor obiecte într-un singur obiect pe baza caracteristicilor comune), eşantionarea (poate fi aleatoare, prin stratificare, adaptivă), reducerea dimensionalităţii (reducerea numărului de atribute prin tehnici precum Analiza Componentelor Principale, Descompunerea Valorilor Singulare, Analiza factorială, Scalare multidimensională, etc), selectarea atributelor importante pentru studiul de data mining, crearea unor noi atribute, discretizarea sau binarizarea (anumiţi algoritmi nu pot lucra decât cu date nominale, deci datele numerice trebuie discretizate), transformarea atributelor (se poate face prin funcţii matematice, normalizare, etc). Capitolul prezintă detaliat aspecte legate de tehnicile de preprocesare menţionate. O analiză statistică realizată la nivel de atribut poate extrage unele informaţii interesante din date şi ajută la înţelegerea datelor. De multe ori datele trebuie convertite din formatul în care ele sunt stocate (Microsoft Excel, Microsoft Access, Microsoft Fox Pro, Oracle, MySQL, etc), în formatul cerut de instrumentul de data mining utilizat pentru analiza datelor. În continuare sunt prezentate aspecte legate de preprocesarea datelor cu Weka, modul în care Weka se poate conecta la baze de date şi a fost îmbunătăţită tehnologia de preprocesare prin conceperea şi implementarea unui instrument software care permite conectarea foarte uşoară la mai multe formate de baze de date, afişarea conţinutului, structurii bazei de date, a relaţiilor dintre tabele şi preluarea datelor din mai multe tabele şi exportarea lor în format arff fără a fi nevoie de cunoştinţe de SQL. Validarea instrumentului propus (numit Arff Convertor) s-a realizat preluând cu ajutorul lui, datele despre studenţii de la facultatea de Automatică şi Calculatoare din anul universitar 2008-2009, transformând datele din două tabele dbf relaţionate care conţin date civice şi date şcolare despre studenţi în format arff şi apoi încărcând în Weka aceste date şi realizând o analiză statistică la nivel de atribut.

Următoarea etapă de după preprocesarea datelor este construirea şi testarea modelului (modelelor).

Modelele se construiesc cu ajutorul unor algoritmi specializaţi. În literatura de specialitate există un număr foarte mare de algoritmi de clasificare şi numărul acestora continuă să crească, pentru că nu s-a descoperit până în prezent vreun algoritm care să fie mai bun decât ceilalţi pe orice set de date şi nu s-a descoperit încă vreun algoritm care să asigure o acurateţe a predicţiei de 100 % pe orice set de date. Câţiva algoritmi clasici folosiţi pentru clasificarea datelor, cunoscuţi în literatură pentru performanţele lor sunt Naive Bayes, k-Nearest Neighbor şi C4.5. Principiul de funcţionare al acestora este prezentat în introducere. Două categorii de algoritmi moderni utilizaţi la clasificarea datelor sunt algoritmii de clasificare bazaţi pe algoritmul Ant Colony Optimization şi algoritmii de clasificare bazaţi pe algoritmi genetici. Aspecte legate de aceste categorii de algoritmi precum şi contribuţiile aduse la clasificarea datelor cu fiecare din aceste două categorii de algoritmi sunt prezentate în capitolele 3 şi 4.

Algoritmul Ant Colony Optimization a fost propus de către Dorigo et al. în anul 1996 pentru optimizarea funcţiilor şi de către Parpineli et al. în anul 2001 pentru clasificarea datelor. În anul 2007 algoritmul a fost implementat într-un sistem Open-Source realizat în

Java, numit Ant Miner şi distribuit pe Internet. De când a fost propus şi până în prezent au fost realizate multe studii cu ajutorul acestui algoritm şi foarte mulţi cercetători l-au studiat şi au propus diverse îmbunătăţiri. De asemenea problema analizei comportamentului algoritmului la diferite valori ale parametrilor de intrare a fost lăsată mereu ca o direcţie de cercetare.

În capitolul 3 este prezentată în detaliu funcţionarea algoritmului ACO, atât pe baza studiului unor articole în care acesta este prezentat cât şi pe baza studiului codului algoritmului. A fost analizat comportamentul algoritmului rulând 64 de configuraţii de parametri pe fiecare din cele 10 seturi de date alese pentru analiză. S-au analizat acurateţea predicţiei şi timpul de execuţie. Detalii privind aceste seturi de date au fost prezentate în introducere. Rezultatele obţinute în urma rulărilor algoritmului ACO pe aceste seturi de date au fost analizate atât empiric (vizual, comparativ), cât şi matematic construind modele de regresie şi matrice de corelare şi analizând coeficienţii de regresie şi pe cei din matricele de corelare. Rezultatul analizei făcute a fost emiterea unei liste de recomandări privind alegerea parametrilor algoritmului.

În continuare a fost realizată o sinteză privind îmbunătăţirile care au fost aduse algoritmului de către diverşi cercetători. Acestea vizează: modul de calcul al funcţiei euristice, modul de calcul al cantităţii de feromon, modul de calcul al acurateţei predicţiei, utilizarea mai multor tipuri de feromon, câte unul pentru fiecare clasă, utilizarea unui intensificator al calităţii pentru a premia regulile bune şi a le penaliza pe cele slabe, s-a permis utilizarea operatorului de negare logică, au fost introduse furnici încăpăţânate care ţin cont de propria istorie când adaugă termeni la regulă, au fost introduse furnici cu personalitate, unele dau mai multă importanţă cantităţii de feromon, altele funcţiei euristice.

Au fost propuse 4 îmbunătăţiri pentru algoritm. Acestea vizează modul de calcul al cantităţii iniţiale de feromon a termenilor – s-a demonstrat că înlocuirea formulei de calcul al feromonului iniţial cu atribuirea unei valori constante pentru feromonul iniţial duce la rezultate mai bune, modul de actualizare a feromonului – a fost introdus un factor de calitate configurabil de către utilizator care permite premierea regulilor bune şi penalizarea celor slabe dar şi păstrarea formulei iniţiale din Ant Miner atunci când utilizatorul alege valoarea 1, modul de evaluare a calităţii regulilor – a fost propus un mod de evaluare a regulilor care ţine cont de acurateţea regulii dar şi de comprehensibilitatea regulilor descoperite şi modul de selecţie a termenilor – a fost introdus un factor de convergenţă (configurabil de către utilizator) care determină în ce măsură se vor alege termenii care au probabilitatea cea mai mare de a fi aleşi (convergenţă rapidă) sau vor fi aleşi termeni după principiul ruletei, proporţional cu performanţa, lucru care determină mai multă explorare.

Îmbunătăţirile au fost operate în sistemul Open Source numit Ant–Miner şi pentru a fi deosebit de sistemul original a fost numit Ant-r-Miner. Atât algoritmul original cât şi cel propus a fost rulat cu aceleaşi valori ale parametrilor de intrare pe cele 10 seturi de date şi rezultatele obţinute au fost comparate din punct de vedere al acurateţei predicţiei şi din punct de vedere al timpului de execuţie. Algoritmul propus a obţinut rezultate mai bune pe 7 seturi de date din 10, în ceea ce priveşte acurateţea predicţiei şi timpul de execuţie, deci are un potenţial mare de aplicare pe noi seturi de date.

În capitolul 4 este tratată în continuare problema algoritmilor utilizaţi pentru a construi modele clasificatoare, de data aceasta atenţia fiind concentrată asupra algoritmilor genetici. Sunt prezentate aspecte generale legate de algoritmii genetici, a fost realizată o sinteză privind algoritmii genetici propuşi la clasificarea datelor şi a fost propus un nou algoritm genetic, numit AGR, care diferă de cei propuşi în literatură prin următoarele:

- permite recalcularea clasei în urma încrucişării
- permite stabilirea unui procent minim de acoperire pentru regulile descoperite, rularea repetată a algoritmului până se atinge acest procent

- permite adăugarea unei reguli implicite listei regulilor descoperite
- permite adăugarea la lista de reguli descoperite doar a celei mai bune reguli obţinută după rularea algoritmului pentru numărul de generaţii stabilit sau a tuturor regulilor descoperite
- utilizează o funcţie *fitness* diferită de cele utilizate de alţi algoritmi

Algoritmul genetic propus a fost implementat în Weka, contribuind în acest fel la îmbunătăţirea tehnologiei de clasificare.

A fost analizat comportamentul algoritmului rulând 14 configuraţii de parametri pe 5 seturi de date reale. S-a analizat comportamentul algoritmului dacă se utilizează toate regulile descoperite de fiecare iteraţie, doar cea mai bună din regulile descoperite de fiecare iteraţie şi dacă se utilizează diferite funcţii fitness. La fiecare rulare au fost folosite pentru antrenare 80 % din instanţe şi pentru testare 20 % din instanţe. Fiecare configuraţie de parametri a fost rulată de 5 ori pe fiecare set de date şi ceea ce se prezintă în lucrare sunt valori medii. Analiza a urmărit parametrii de ieşire acurateţea pe mulţimea de antrenament, acurateţea pe mulţimea de test, acoperirea pe mulţimea de test, acoperirea pe mulţimea de antrenament, numărul de reguli descoperite, timpul de execuţie, numărul regulilor descoperite şi numărul de condiţii din regulile descoperite. În urma analizei a fost emis un set de recomandări privind alegerea parametrilor. Rezultatele obţinute de AGR au fost comparate cu cele obţinute de algoritmii Naive Bayes, k-Nearest Neighbor şi J4.8. În urma comparaţiei AGR a obţinut rezultate mai bune decât algoritmii menţionaţi pe 4 seturi de date din 5, în ceea ce priveşte acurateţea predicţiei, punctul lui slab fiind timpul de execuţie.

Următorul pas al oricărui studiu de clasificare a datelor- după ce acestea au fost preprocesate şi au fost construiţi şi testaţi clasificatori- este de a utiliza aceşti clasificatori pentru a realiza predicţii.

Capitolul 5 prezintă aspecte legate de realizarea predicţiilor cu Weka. Este prezentat critic modul în care se pot realiza predicţii în Weka şi este demonstrat că soluţia propusă de Weka lasă mult de dorit.

A fost propusă o nouă soluţie pentru realizarea predicţiilor, interfaţa principală a lui Weka a fost transformată într-o interfaţă dinamică care depinde de setul de date care este deschis pentru analiză. Interfaţa va conţine un număr de label-uri egal cu numărul de atribute al setului de date pentru fiecare atribut câte un label cu denumirea atributelor. Pentru atributele nominale în interfaţă sunt afişate combobox-uri din care utilizatorul poate alege una din valorile nominale posibile, iar pentru atributele numerice în interfaţă sunt afişate text boxuri în care utilizatorul poate introduce valorile numerice. După completarea valorilor instanţei care urmează să fie prezisă se apasă comanda *Predict* şi clasa prezisă este afişată lângă buton. Gradul de încredere în predicţia realizată este afişat sub forma unui grafic nou introdus în interfaţă care redă acurateţea predicţiei.

În capitulul 6 au fost utilizate îmbunătăţirile tehnologice propuse în toate capitolele precedente pentru a analiza datele despre naşterile care au avut loc în anul 2010 la *Clinica de Obstetrică - Ginecologie Bega, Timişoara.* Au fost analizate date despre 2326 de naşteri. Etapele studiului au fost următoarele:

- Preprocesarea în Microsoft Excel a datelor legate de naşterile din 2010 prin analiza fiecărei coloane, eliminarea instanţelor cu valori lipsă sau valori introduse greşit, corectarea valorilor introduse greşit, eliminarea atributelor neimportante pentru analiză, crearea unor noi atribute, etc
- Transformarea datelor din format *xls* în formatul *arff*, specific *Weka* cu ajutorul aplicaţiei *Arff Convertor*, concepută, implementată şi prezentată de autor în cadrul capitolului 2
- Preprocesarea datelor cu ajutorul lui *Weka*, prin discreditarea atributelor numerice

- Analiza statistică la nivel de atribut, care a condus la obţinerea unor informaţii noi şi interesante privind naşterile
- Dezvoltarea unor modele de clasificare cu algoritmii *Naive Bayes, J48, k-Nearest-Neighbour* din *Weka*, algoritmi care au fost prezentaţi în introducere
- Dezvoltarea unor modele de clasificare cu algoritmii *Ant-Miner şi Ant-r-Miner*, prezentaţi în capitolul 3
- Dezvoltarea unui model de clasificare care poate estima cu o precizie de aprox. 85 % unul din cele 5 intervale în care indicele apgar al nou-născutului va lua valoare. Modelul a fost construit cu ajutorul algoritmului AGR, care a fost conceput, şi implementat şi prezentat de autor în cadrul capitolului 4
- Realizarea de predicţii – utilizarea interfeţei dinamice a lui *Weka,* concepută şi prezentată de autor în cadrul capitolului 5 pentru realizarea predicţiilor
- Setul de reguli descoperit a fost transpus în format XML cu ajutorul unei aplicaţii concepute şi implementate de autor, pentru a servi ca intrare unui sistem care le transpune în A*rden Syntax,* cu scopul de a utiliza regulile descoperite în sisteme de tip *Clinical Decision Suport* pentru generarea de recomandări medicale

Valoarea medicală a modelului clasificator este următoarea: Indicele apgar reprezintă o apreciere a funcţiilor vitale şi a capacităţii de adaptare a fătului la condiţiile din mediul extrauterin. Este de dorit ca valoarea acestui indicator să fie cât mai mare pentru fiecare nou-născut. Modelul construit poate fi utilizat înainte de naştere, pentru a realiza predicţii ale intervalului în care se va plasa acest indice (unul din cele 5 intervale) în funcţie de datele care se cunosc despre mamă, copil şi intervenţiile medicale care urmează să se realizeze pentru a ajuta naşterea. Dacă rezultatele predicţiei dau un indice apgar mic este necesar să se realizeze mai multe predicţii cu modelul construit, la fiecare predicţie modificând parametrii de intrare care pot fi modificaţi în practică de medici, pentru ca acest indice apgar să se maximizeze (de exemplu se poate modifica tipul de naştere, sau se poate ajuta fătul să ajungă în poziţia normală, etc). Ceea ce se caută este modul în care pot fi modificate intrările astfel încât la ieşire să se obţină o valoare mulţumitoare. Un exemplu al utilizării modelului este: se realizează predicţii şi se analizează în ce interval ar primi nou născutul nota (indicele apgar) dacă fătul s-ar naşte natural respectiv prin cezariană. Dacă se estimează că ar primi notă mai mare dacă s-ar naşte natural, atunci poate că mamele şi / sau medicii se vor gândi de două ori înainte de a face cezariană. Situaţia poate fi valabilă şi invers.

Într-o formulare şi prezentare sintetică, principalele contribuţii aduse prin teza de doctorat sunt următoarele:

1. **o sinteză originală privind domeniul data mining**:
   - etapele procesului de data mining
   - algoritmii utilizaţi la clasificarea datelor
   - domeniile în care tehnicile de *data mining* se aplică cu succes
   - instrumentele utilizate în studii de *data mining*
   - tehnicile de preprocesare a datelor
   - algoritmul ACO şi îmbunătăţirile aduse acestuia în literatura de specialitate
   - algoritmii genetici utilizaţi la clasificarea datelor
2. **o sinteză originală privind principalele tehnici de preprocesare a datelor**, bazată pe studiul a 30 de referinţe, în cadrul căreia au fost tratate: tehnicile de preprocesare, agregarea, eşantionarea, reducerea dimensionalităţii (prin analiza componentelor principale, descompunerea valorilor singulare, analiză factorială, încapsularea locală lineară, scalare multidimensională), selectarea de submulţimi

de caracteristici, crearea caracteristicilor, discretizarea şi binarizarea, transformarea variabilelor, etc.

3. **o serie de îmbunătăţiri ale tehnologiei de preprocesare a datelor** prin concepţia şi dezvoltarea unui instrument software care permite conversia datelor din 5 formate de baze de date în formatul de date folosit de *Weka* (arff); instrumentul propus dispune de interfeţe dinamice, în acord cu datele care se prelucrează, iar utilizarea lui este simplă şi nu necesită realizarea de configurări în sistemul de operare, în *Weka* sau cunoştinţe de SQL
4. **o analiză** a influenţei pe care o au parametrii de intrare ai algoritmului ACO asupra acurateţei predicţiei şi asupra timpului de execuţie şi **o listă** de recomandări cu privire la alegerea acestora
5. **o serie de îmbunătăţiri aduse algoritmului ACO**, care vizează modul de calcul al cantităţii iniţiale de feromon a termenilor, modul de actualizare a feromonului, modul de evaluare a calităţii regulilor descoperite şi modul de selecţie a termenilor, confirmate prin teste şi analize comparate bazate pe date reale, însoţite de implementări ale lor, sintetizate sub denumirea Ant-r-Miner
6. **un nou algoritm genetic** – AGR-, validat prin teste comparative, algoritm care, după cunoştinţele autorului, diferă de ceilalţi algoritmi genetici de clasificare prin următoarele:
   - permite recalcularea clasei în urma încrucişării
   - permite stabilirea unui procent minim de acoperire pentru regulile descoperite şi rularea repetată a algoritmului până se atinge acest procent
   - permite adăugarea unei reguli implicite listei regulilor descoperite
   - permite adăugarea la lista de reguli descoperite doar a celei mai bune reguli obţinută după rularea algoritmului pentru numărul de generaţii stabilit sau a tuturor regulilor descoperite
   - utilizează o funcţie *fitness* diferită de cele utilizate de alţi algoritmi genetici
7. **o implementare originală integrată în aplicaţia *open-source* Weka a algoritmului AGR**, ce extinde capabilităţile Weka de clasificare a datelor
8. **concepţia şi implementarea unei noi metode de a realiza predicţii în aplicaţia *open-source* Weka, cu ajutorul unei interfeţe dinamice** ce conţine controale în acord cu setul de date analizat şi care uşurează şi clarifică mult procesul de realizare a predicţiilor
9. **concepţia şi implementarea unor soluţii de transformare a interfeţei de clasificare a datelor din aplicaţia *open source* Weka** într-o interfaţă mult mai simplă, inteligibilă şi prietenoasă, prin vizualizarea sub formă de grafic 3d a celei mai importante măsuri (acurateţea predicţiei) obţinute în urma testării modelului şi prin afişarea informaţiilor din secţiunea *Classifier output* (a indicatorilor ”complicaţi”) într-o nouă fereastră *Advanced Information*
10. **concepţia de modele clasificatoare privind naşterile**, atât cu algoritmii clasici *Naive Bayes, J48, k-Nearest-Neighbour, Ant-Miner*, dar şi cu algoritmii propuşi: *Ant-r-Miner* şi *AGR*
11. **o analiză comparativă a algoritmilor clasici şi a algoritmului propus, *AGR***, care evidenţiază că cele mai bune performanţe se obţin cu algoritmul *AGR*, care permite construirea unui clasificator ce poate estima cu o precizie de aprox. 85 % unul din cele 5 intervale în care indicele apgar al nou-născutului va lua valoare
12. **realizarea de predicţii** utilizând interfaţa dinamică a lui *Weka,* concepută şi implementată de autor

13. **concepţia şi implementarea unei aplicaţii pentru transpunerea regulilor descoperite în format XML**, pentru a servi ca intrare în sistemele de tip *Clinical Decision Support* pentru generarea de recomandări medicale

În continuare sunt prezentate contribuţiile aduse de teza de doctorat, în totalitatea lor, grupate pe capitole:

*În **capitolul 1**, aflat pe post de introducere, contribuţiile sunt:*

- conturarea domeniului, cu sublinierea importanţei lui
- prezentarea sistematică a termenilor domeniului şi a principalelor definiţii
- propunerea unei noi definiţii
- prezentarea originală a etapelor procesului de data mining, pe baza experienţei practice a autorului
- realizarea unei sinteze ample asupra aplicării tehnicilor de data mining în domeniile e-commerce, educaţie, medicină, combaterea terorismului
- realizarea unei sinteze asupra sistemelor de *data mining* utilizate de agenţiile federale din SUA, cu scopul de sublinia puterea şi importanţa tehnologiei *data mining*, precum şi nivelul de vârf la care s-a ajuns cu utilizarea ei
- prezentarea rezumativă şi sistematică, într-o abordare proprie, a aspectelor generale legate de clasificarea datelor, precum şi a principiilor de bază privind funcţionarea a trei algoritmi de clasificare de referinţă (*Naive Bayes, k-Nearest Neighbor, C4.5*) care au fost utilizaţi în studiile realizate în capitolele 4, 5 şi 6
- prezentarea rezumativă a aspectelor caracteristice a două dintre cele mai populare instrumente, la nivel mondial, utilizate pentru a realiza studii de data mining (*Weka* şi *R*), cu indicarea avantajelor şi dezavantajelor pe care le are fiecare instrument, printr-o analiză comparată
- prezentarea aspectelor caracteristice privitoare la cele 10 seturi de date reale, alese de autor de pe *UCI Machine Learning Repository*, care au fost utilizate pe parcursul tezei, în capitolele 3, 4 şi 5 pentru a testa şi valida algoritmii propuşi şi îmbunătăţirile tehnologice aduse

*În **capitolul 2**, dedicat aspectelor legate de preprocesarea datelor, contribuţiile sunt:*

- realizarea unei sinteze ample, cu privire la principalele tehnici de preprocesare a datelor, bazată pe studiul a 30 de referinţe, în cadrul căreia au fost tratate: tehnicile de preprocesare, agregarea, eşantionarea, reducerea dimensionalităţii (prin analiza componentelor principale, descompunerea valorilor singulare, analiză factorială, încapsularea locală lineară, scalare multidimensională), selectarea de submulţimi de caracteristici, crearea caracteristicilor, discretizarea şi binarizarea, transformarea variabilelor, etc.
- generarea unei serii de îmbunătăţiri ale tehnologiei de preprocesare a datelor prin concepţia şi dezvoltarea unui instrument software care permite conversia datelor din 5 formate de baze de date în formatul de date folosit de *Weka* (arff); instrumentul propus dispune de interfeţe dinamice, în acord cu datele care se prelucrează, iar utilizarea lui este simplă şi nu necesită realizarea de configurări în sistemul de operare, în *Weka* sau cunoştinţe de SQL
- realizarea unei analize statistice a datelor studenţilor de la facultatea de Automatică şi Calculatoare din anul universitar 2008-2009 cu ajutorul lui *Arff Convertor* şi *Weka*, analiză care oferă o imagine de ansamblu asupra provenienţei studenţilor, pregătirii lor, etc, menită să demonstreze utilitatea şi buna funcţionare a lui Arff Convertor

*În **capitolul 3**, dedicat aspector legate de clasificarea datelor cu ajutorul algoritmului Ant Colony Optimization, contribuţiile sunt*:

- prezentarea detaliată a algoritmului, într-o manieră originală, bazată atât pe documentaţia existentă, cât şi pe studiul codului algoritmului
- generarea unui studiu original asupra parametrilor algoritmului, prin rularea a câte 64 de configuraţii pe fiecare din cele 10 seturi de date alese şi analizarea rezultatelor prin comparaţie vizuală, respectiv matematic, construind şi analizând modele de regresie şi matrice de corelare
- generarea unei liste de recomandări privind modul de alegere a parametrilor, utile pentru utilizatorii acestui algoritm
- generarea unei sinteze originale a îmbunătăţirilor aduse algoritmului din anul 2001 şi până în prezent
- generarea unei serii de îmbunătăţiri ale algoritmului, care vizează modul de calcul al cantităţii iniţiale de feromon a termenilor, modul de actualizare a feromonului, modul de evaluare a calităţii regulilor descoperite şi modul de selecţie a termenilor; modificările propuse au fost implementate în aplicaţia open source *Ant Miner*. Aplicaţia care conţine algoritmul îmbunătăţit a fost numită *Ant-r-Miner*. S-a demonstrat că îmbunătăţirile aduse conduc la o mai bună performanţă utilizând date reale.

*În **capitolul 4**, dedicat prezentării unui nou algoritm genetic pentru clasificarea datelor, denumit AGR, contribuţiile sunt:*

- prezentarea comprehensivă a structurii generale a unui algoritm genetic
- realizarea unei sinteze originale cu privire la algoritmii genetici utilizaţi la clasificarea datelor
- propunerea unui nou algoritm genetic pentru clasificarea datelor, care, după cunoştinţele autorului, diferă de ceilalţi algoritmi genetici de clasificare prin următoarele:
    - permite recalcularea clasei în urma încrucişării
    - permite stabilirea unui procent minim de acoperire pentru regulile descoperite şi rularea repetată a algoritmului până se atinge acest procent
    - permite adăugarea unei reguli implicite listei regulilor descoperite
    - permite adăugarea la lista de reguli descoperite doar a celei mai bune reguli obţinută după rularea algoritmului pentru numărul de generaţii stabilit sau a tuturor regulilor descoperite
    - utilizează o funcţie *fitness* diferită de cele utilizate de alţi algoritmi genetici
- îmbunătăţirea tehnologiei de clasificare a datelor prin implementarea lui *AGR* într-un instrument *open source* utilizat pe scară largă (*Weka*)
- analiza originală a comportamentului algoritmului la diferite valori ale parametrilor de intrare şi la utilizarea a diferite funcţii *fitness* şi emiterea unui set de recomandări cu privire la alegerea acestora

În ***capitolului 5***, dedicat tratării problemei realizării de predicţii cu ajutorul unui model construit şi testat pe *Weka, contribuţiile sunt*:

- prezentarea detaliată, într-o manieră originală, a modului în care *Weka* poate fi utilizat pentru a realiza predicţii
- concepţia şi implementarea unei interfeţe dinamice care conţine controale în acord cu setul de date analizat şi care uşurează şi clarifică mult procesul de realizare a predicţiilor

- prezentarea sintetică a mărimilor de ieşire furnizate de *Weka* în urma testării modelului
- concepţia şi implementarea unor soluţii de transformare a interfeţei de clasificare a datelor într-o interfaţă mult mai simplă, inteligibilă şi prietenoasă, prin vizualizarea sub formă de grafic 3d a celei mai importante măsuri (acurateţea predicţiei) obţinute în urma testării modelului şi prin afişarea informaţiilor din secţiunea *Classifier output* (a indicatorilor "complicaţi") într-o nouă fereastră *Advanced Information*
- testarea pe 3 seturi de date medicale a contribuţiilor şi validarea lor, prin rezultatele foarte bune obţinute

*În **capitolul 6**, dedicat utilizării tehnologiei prezentate şi a îmbunătăţirilor tehnologice propuse în capitolele precedente, pentru a analiza şi clasifica datele despre cele 2326 naşteri care au avut loc la Clinica de Obstetrică - Ginecologie Bega, Timişoara în anul 2010, contribuţiile sunt:*

- preprocesarea datelor în Microsoft Excel, Arff Convertor şi Weka
- realizarea unei analize statistice la nivel de atribut, care a condus la obţinerea unor informaţii interesante, cum ar fi: greutatea medie a nou născuţilor este 3182 g, au născut natural 925 de femei (40 %) şi prin cezariană 1352 de femei (60%), epiziotomia a fost necesară in 70 % din naşterile naturale, etc
- realizarea de modele clasificatoare, atât cu algoritmii clasici *Naive Bayes, J48, k-Nearest-Neighbour, Ant-Miner*, dar şi cu algoritmii propuşi: *Ant-r-Miner* şi *AGR*
- realizarea unei analize comparate a algoritmilor clasici şi a algoritmului propus, *AGR*, care evidenţiază că cele mai bune performanţe se obţin cu algoritmul *AGR*, care permite construirea unui clasificator ce poate estima cu o precizie de aprox. 85 % unul din cele 5 intervale în care indicele apgar al nou-născutului va lua valoare
- realizarea de predicţii utilizând interfaţa dinamică a lui *Weka,* concepută şi prezentată de autor în cadrul capitolului 5 pentru realizarea predicţiilor
- transpunerea regulilor descoperite în format XML, cu ajutorul unei aplicaţii concepute şi implementate de autor pentru a servi ca intrare în sistemele de tip *Clinical Decision Support* pentru generarea de recomandări medicale

Direcţiile de dezvoltare pe care teza le lasă deschise vizează, pe de o parte, continuarea îmbunătăţirii instrumentelor software care implementează algoritmii propuşi Ant-r-Miner şi Weka cu AGR în aşa fel încât utilizatorul să-şi poată alege şi / sau introduce funcţia fitness pe care doreşte să o folosească, pe de altă parte, implementarea mai multor metode de selecţie în AGR, iar în al treilea rând, aplicarea tehnologiei propuse în domeniul medical. O altă direcţie de urmat va putea fi concepţia unei modalităţi de utilizare a algoritmului albină în problema clasificării datelor.