UNIVERSITATEA DIN CRAIOVA
Facultatea de Automatică, Calculatoare şi Electronică

Laviniu Aurelian Bădulescu

- REZUMATUL TEZEI DE DOCTORAT-

OPTIMIZAREA UNOR ALGORITMI CU ARBORI DE DECIZIE ÎN „DATA MINING”

Conducător ştiinţific: Acad. Prof. Univ. Dr. Ing. Mircea Petrescu

1. CUPRINS

2. CUVINTE CHEIE: *data mining*, *problema clasificării, arbori de decizie*, *setul de date de antrenare, setul de date de test, clasă, etichete ale clasei, criterii de selecție a atributului*, *extragerea cunoştințelor, reguli de decizie*, *arbore de decizie neredus, metode de reducere, arbore de decizie redus, rata erorii de clasificare*, *precizia clasificării*.

3. SINTEZE ALE PĂRȚILOR PRINCIPALE ALE TEZEI DE DOCTORAT

Obiectivele tezei de doctorat au vizat:

- optimizarea algoritmilor cu arbori de decizie (AD) în direcția găsirii unor criterii optime de divizare ale setului de date de la nivelul unui nod al AD şi a unor metode optime de reducere a AD;
- dezvoltarea unui cadru algoritmic unificat pentru abordarea algoritmilor cu AD, a criteriilor de divizare şi a metodelor de reducere;

- dezvoltarea unui cadru experimental în care, pe baza unui sistem propus de programe pentru inducerea, reducerea şi executarea AD, a generării matricei de confuzie şi a regulilor de decizie asociate AD, să fie realizata cercetarea.

Teza de doctorat este structurată pe şase capitole precedate de un cuprins al tezei şi urmate de bibliografia utilizată la scrierea acesteia. Cuprinsul este structurat astfel încât, în problematica domeniului, include părţile: Introducere (Capitolul 1), Stadiul cunoaşterii (Capitolul 2), Contribuţii proprii (Capitolul 3, Capitolul 4 şi Capitolul 5), Concluzii finale (Capitolul 6) şi Bibliografie.

În **Capitolul 1**, **Introducere**, se realizează introducerea în problematica temei tratate. Astfel, se prezintă raportul dintre extragerea de cunoştinţe din bazele de date şi data mining (DM), problema clasificării, conceptele fundamentale privind AD, precum şi avantajele şi dezavantajele algoritmilor cu AD. În finalul capitolului se prezintă modul cum este organizată teza şi care sunt principalele contribuţii originale ale autorului ei. Una dintre principalele contribuţii personale la nivel teoretic, este construirea unui cadru unificat în care au fost tratate unitar câteva zeci de criterii de divizare a setului de date de la nivelul unui nod al AD şi, în acelaşi cadru matematic, prezentarea şi a câtorva metode de reducere a AD.

Considerând că principalele abordări de optimizare a algoritmilor cu AD trebuie să meargă în primul rând pe două direcţii: ale alegerii unui criteriu de divizare optim şi a celei mai bune metode de reducere, contribuţia personală majoră a tezei la nivel practic, este reprezentată de descoperirea, pe baza unui număr foarte mare de experimente personale, a unui criteriu de divizare care manifestă sistematic cele mai bune performanţe, comparativ cu câteva zeci de alte criterii luate în consideraţie în experimente. În acelaşi timp, s-a demonstrat superioritatea metodei de reducere a AD cu nivel de încredere. Astfel au fost concepuţi noi algoritmi de AD, optimizaţi, reprezentând o contribuţie personală a tezei.

O altă contribuţie majoră a tezei de doctorat a fost reprezentată de dezvoltarea unui cadru experimental în care, pe baza unui sistem de programe pentru inducerea, reducerea şi executarea AD, a generării matricei de confuzie şi a regulilor de decizie asociate AD, au fost realizate numeroase experimente folosind baze de date cu caracteristici foarte diferite. Rezultatele experimentale au fost comparate cu numeroase rezultate ale altor cercetători care au experimentat pe aceleaşi baze de date, dar folosind alte tipuri de algoritmi de DM, constatându-se că rezultatele experimentelor personale sunt la fel de bune ca cele mai bune valori din literatură.

O altă contribuţie personală importantă este reprezentată de studiul din Capitolul 2, rezultatul parcurgerii exhaustive şi la zi a unui volum imens de lucrări în domeniul AD, coroborat cu un efort susţinut de sintetizare şi sistematizare a materialului lecturat. Acest capitol prezintă o sinteză personală a principalilor algoritmi de AD, realizată pe baza unei bibliografii exclusiv din literatura străină.

Capitolul 2, **Prezentarea literaturii existente asupra arborilor de decizie**, realizează o sinteză personală a principalilor algoritmi de AD, obţinută în urma parcurgerii şi sistematizării unei importante bibliografii din literatura străină. Astfel, sunt prezentaţi numeroşi algoritmi de AD: algoritmii AID, THAID, CHAID şi FIRM, algoritmul CART, ID3, C4.5 şi alţi algoritmi înrudiţi, AD leneşi, algoritmii cu AD din pachetul SAS, algoritmi cu AD multidimensionali, algoritmi cu AD multidimensionali oblici, algoritmi cu AD omnidimensionali, algoritmi cu AD uituci, algoritmi cu AD paraleli, algoritmii cu AD din pachetul WEKA, AD cu algoritmi genetici, algoritmi cu AD pentru fluxuri de date, algoritmi pentru AD ortogonali, AD hibrizi, AD adaptivi online, algoritmi cu AD distribuiţi precum şi alţi algoritmi cu AD.

Capitolul 3, Algoritmi de inducere şi reducere a arborilor de decizie şi criterii de selecţie a atributului folosite în experimente, reprezintă o contribuţie originală, propunând un cadru algoritmic unificat pentru abordarea algoritmilor de AD şi furnizează o descriere aprofundată a

zeci de măsuri de selecție a atributului pentru divizarea setului de date corespunzător unui nod şi a câtorva metode de reducere a AD (reducerea complexității costului, reducerea cu scăderea erorii, reducerea cu eroare minimă, reducerea pesimistă, reducerea pe baza erorii/reducerea cu nivel de încredere, reducerea optimă şi reducerea pe baza principiului lungimii minime a descrierii). Capitolul 3 prezintă algoritmii propuşi pentru inducerea şi reducerea AD, criterii bazate pe impuritate, criterii bazate pe normalizarea impurității, criterii binare, K2 şi Bayes-Dirichlet utilizate ulterior în experimente. Totodată se prezintă modalitatea de transformare a AD în setul regulilor de decizie şi modul de tratare a atributelor cu valori lipsă.

Capitolul 4, Materialele şi metoda de lucru folosite în experimente, prezintă cele opt baze de date folosite în experimente, ale căror dimensiuni ajung până la aproape 600.000 de cazuri, 55 de atribute (continui şi discrete), clase cu 7 etichete, numeroase cazuri cu valori lipsă şi înregistrări duplicate sau contradictorii. Tot în acest capitol se prezintă sistemul de programe propus şi utilizat în experimente: programul pentru inducerea AD, cel pentru reducerea AD, cel pentru testarea preciziei de clasificare a AD, cel pentru calcularea matricei de confuzie şi cel pentru generarea regulilor de decizie asociate AD.

Capitolul 5, **Rezultate experimentale şi discuții**, prezintă experimentele realizate pentru fiecare din cele 8 baze de date pe baza unui set identic de prelucrări. Performanțele fiecărui pas al prelucrărilor au fost prezentate, în tabele şi diagrame însoțite de comentarii, la fiecare pas al experimentelor.

1. Primul pas al experimentelor a fost reprezentat de inducerea AD pe baza setului de date de antrenare. La acest pas au fost induşi 28 de AD, corespunzător celor 28 de măsuri de selecție a atributului. Au fost luate în considerare valorile unor parametrii ai procesului de construcție al AD: numărul de noduri al AD indus, numărul de atribute necesar inducerii AD, numărul de nivele ale AD, timpul de creştere a AD, mărimea fişierului conținând AD indus. S-au comparat şi discutat aceste valori, s-au calculat indicatori statistici, s-au prezentat tabele şi diagrame însoțite de comentarii, extrăgându-se concluzii asupra comportamentului diferit al celor 28 de AD induşi cu cele 28 de măsuri.

2. La al doilea pas al experimentelor, aceşti 28 de AD în formă neredusă au fost prelucrați pentru a se extrage din ei cele 28 de seturi de reguli de decizie, câte un set de reguli de decizie pentru fiecare AD obținut la pasul întâi. Au fost luate în considerare valorile unor parametrii ai procesului de extragere a regulilor de decizie din AD: numărul regulilor de decizie, timpul de construire a fişierului conținând regulile de decizie, timpul de citire a fişierului conținând AD, mărimea fişierului conținând setul de reguli de decizie. S-au comparat şi discutat aceste valori, s-au calculat indicatori statistici, s-au prezentat tabele şi diagrame însoțite de comentarii, extrăgându-se concluzii asupra comportamentului diferit al celor 28 de AD induşi cu cele 28 de măsuri.

3. Al treilea pas al experimentelor, realizat pe fiecare bază de date, a presupus execuția celor 28 de AD induşi la pasul întâi. Aceste experimente au verificat modelul de clasificare reprezentat de AD pe baza unui set de date de test, set de date necunoscut la pasul întâi, când a fost generat AD. Este cel mai important pas al experimentelor, deoarece acum se obține valoarea ratei erorii de clasificare pe datele de test sau precizia modelului. Pentru fiecare bază de date se obțin 28 de valori ale preciziei clasificării pe datele de test a AD în formă neredusă, corespunzător fiecăreia din cele 28 de măsuri cu care a fost indus AD. Prin compararea acestor valori se pot trage concluzii asupra performanțelor uneia sau alteia dintre măsurile de selecție a atributului.

4. Al patrulea pas al experimentelor a vizat reducerea celor 28 de AD, obținuți la pasul întâi, cu metoda de reducere cu nivel de încredere. Au fost luate în considerare valorile unor parametrii ai procesului de reducere a AD: nivelul de încredere, numărul de atribute necesar construirii AD redus, numărul de noduri şi numărul de nivele ale AD redus, timpul de reducere a fişierului conținând AD neredus, mărimea fişierului conținând AD redus. S-au comparat şi discutat aceste

valori, s-au calculat indicatori statistici, s-au prezentat tabele şi diagrame însoțite de comentarii, extrăgându-se concluzii asupra comportamentului diferit al celor 28 de AD reduşi cu nivel de încredere.

5. Al cincilea pas al experimentelor a constat în extragerea regulilor de decizie din cei 28 de AD reduşi cu metoda de reducere cu nivel de încredere la pasul patru. Prelucrările au semănat cu cele de la pasul 2.

6. Pasul al şaselea a presupus execuția pe datele de test a celor 28 de AD reduşi cu metoda de reducere cu nivel de încredere la pasul patru. Au fost reținute valorile ratei erorii de clasificare pe datele de test, valori care ulterior au fost comparate cu valorile obținute la execuția pe datele de test a AD nereduşi (vezi pasul 3) şi cu valorile obținute la execuția pe datele de test a AD reduşi cu metoda de reducere pesimistă (vezi pasul 7).

7. Al şaptelea pas al experimentelor a vizat reducerea celor 28 de AD, obținuți la pasul 1, cu metoda de reducere pesimistă. Prelucrările au semănat cu cele de la pasul 4.

8. Al optulea pas al experimentelor a constat în extragerea regulilor de decizie din cei 28 de AD reduşi cu metoda de reducere pesimistă la pasul şapte. Prelucrările au semănat cu cele de la pasul 2.

9. Pasul al nouălea a presupus execuția pe datele de test a celor 28 de AD reduşi cu metoda de reducere pesimistă la pasul şapte. Prelucrările au semănat cu cele de la paşii 3 şi 6.

10. La pasul al zecelea s-a realizat o discuție asupra valorilor numărului de reguli de decizie şi a valorilor ratei erorii de clasificare pe datele de test pentru cele trei tipuri de AD: neredus, redus cu metoda de reducere cu nivel de încredere şi redus cu metoda de reducere pesimistă, evidențiind care măsură de selecție a atributului şi care metodă de reducere a avut cel mai bun comportament în experimente.

11. În final, la pasul unsprezece, au fost trecute în revistă rezultatele unor lucrări reprezentative din literatura de specialitate ce prezintă teste de performanță pe baza de date respectivă şi s-au comparat rezultatele obținute în acele experimente cu rezultatele obținute de cele 28 de măsuri şi cele două metode de reducere. Din motive de spațiu, la o singură bază de date (*Image*) am calculat matricele de confuzie.

Toate experimentele pe cele 8 baze de date au fost realizate pe un calculator PC AMD Duron 995 MHz CPU cu 512 MB RAM, sub sistemul de operare Windows XP Professional Version 2002 Service Pack 2, iar codul a fost scris în limbajul de programare C folosind mediul de programare integrat Microsoft Visual C++ 6.0 Enterprise Edition.

În **Capitolul 6,** se prezintă **Concluzii ale experimentelor şi direcții pentru experimente viitoare**. Astfel, folosind trei tehnici se compară performanțele la clasificare ale celor 28 de AD induşi şi reduşi cu două metode de reducere pe cele opt baze de date.

Prima dată comparația se realizează pe baza valorilor mediei aritmetice şi a abaterii standard a valorilor ratei erorii de clasificare pe datele de test pentru toate tipurile de AD, toate bazele de date şi toate măsurile de selecție a atributului. Se constată că cel mai bine clasifică AD indus cu măsura *rciq* (raportul câştigului în informație pătratic), cu o medie a ratei erorii de clasificare pe datele de test de 18,59% (având şi cea mai mică valoare pentru abaterea standard: 14,11), urmată de AD induşi cu măsurile *rel* (relief), *k2*, *bd* (Bayes-Dirichlet), *lmdfa* (lungimea minimă a descrierii în frecvențe absolute) şi *csh* (complexitatea stohastică) cu valori mici pentru abaterea standard, iar cea mai slabă performanță medie o realizează AD indus cu măsura *ciqp* (câştigul în informație pătratic ponderat), cu o valoare medie de 31,12%.

AD indus cu măsura *rciq* are cea mai bună valoare a mediei ratei erorii de clasificare şi cea mai mică abatere standard, deci cea mai mică împrăştiere a valorilor ratei erorii de clasificare pe datele de test în jurul mediei acestora. Acest rezultat ne indică faptul că performanța AD indus cu

măsura *rciq* nu este afectată prea mult de caracteristicile bazei de date. Adică, una din caracteristicile fundamentale ale măsurii *rciq* este independența de domeniu, caracteristică foarte importantă astăzi, când bazele de date sunt formate din date, cu atribute aparținând la domenii diferite, colectate împreună. Diversitatea domeniilor din alcătuirea bazelor de date este unul dintre motivele care au dus la creşterea necesității utilizării unor instrumente de extragere automată a cunoştințelor din bazele de date şi a algoritmilor de învățare inductivă.

Rezultatele experimentelor ne arată că cea mai mică medie a numărului de reguli de decizie îl prezintă AD indus cu măsura *mapd* (media absolută a ponderii dovezii) cu o medie a numărului de reguli de decizie de 1.147,54, iar cea mai slabă performanță medie o realizează AD indus cu măsura *rcs* (raportul câştigului specificității) cu o performanță medie de 1.641,25. Să observăm că AD indus cu măsura *rciq*, care are cea mai bună comportare din punctul de vedere al preciziei clasificării pe datele de test, se plasează pe locul patru în ierarhia celor mai mici medii pentru numărul de reguli de decizie.

A doua modalitate de comparație a valorilor performanțelor de clasificare a celor 28 de AD a fost reprezentată de tehnica *win/tie/loss*. Şi în acest caz, performanțele AD indus cu măsura *rciq* sunt totdeauna cele mai bune. AD indus cu măsura *rciq* prezintă cel mai mare număr de cazuri când are o valoare a performanței mai bună, decât toate celelalte măsuri şi în acelaşi timp, cel mai mic număr de cazuri când are o valoare a performanței mai slabă decât a celorlalte măsuri. Următoarele cinci locuri sunt ocupate de AD induşi cu alte cinci măsuri *csh*, *bd*, *k2*, *lmdfa* şi *rel*, care însă nu mai prezintă proprietatea manifestată de măsura *rciq*, de a avea atât un număr mare de cazuri când depăşesc valorile performanțelor AD induşi cu toate celelalte măsuri, cât şi un număr mic de cazuri când sunt depăşite de valorile performanțelor celorlalți AD.

A treia modalitate de comparație a valorilor performanțelor de clasificare a celor 28 de AD a fost reprezentată de media geometrică a raportului ratei erorii de clasificare. Performanțele AD indus cu măsura *rciq* întrec, în toate cazurile considerate în experimente, performanțele AD induşi cu celelalte 27 măsuri. Îi urmează, în ordine, valorile performanțelor realizate de AD induşi cu măsurile *rel*, *gim* (gini index modificat), *bd*, *k2*, *lmdfa* şi *csh*. Astfel constatăm că, indiferent de criteriul de comparație considerat, optimizarea algoritmilor cu AD prin utilizarea măsurii *rciq* induce AD cei mai performanți. Şi optimizarea pe baza măsurilor *rel*, *k2*, *bd*, *lmdfa* şi *csh* produce AD cu performanțe bune.

Din punctul de vedere al găsirii metodei de reducere optime, analiza valorilor obținute din experimentele prezentate în teză arată că AD redus cu metoda de reducere cu nivel de încredere obține sistematic cele mai bune valori ale preciziei clasificării pe datele de test, simultan cu generarea unui număr mic de reguli de decizie. Pe locul doi se situează performanțele AD redus pesimist, iar pe ultimul loc AD neredus. Să observăm că această diferențiere, care plasează totdeauna AD redus cu nivel de încredere pe primul loc şi AD redus pesimist pe locul doi, este mult mai clară în privința numărului de reguli de decizie, decât în privința acurateței clasificării pe datele de test. În acelaşi timp, să mai observăm că reducerea îmbunătățeşte performanțele AD.

Experimentele pe care dorim să le efectuăm în continuare vor viza baze mult mai mari de date, cu foarte multe atribute şi foarte multe cazuri, pe care dorim să verificăm concluziile pe care le-am obținut pentru bazele de date utilizate până în prezent. În acelaşi timp, vom include în setul de măsuri pe baza cărora se realizează divizarea unui nod şi alte criterii de selecție a atributului şi vom utiliza şi alte metode de reducere a AD, în afara celor două utilizate în experimentele noastre. În acelaşi timp va trebui să ținem cont şi de informațiile furnizate de matricea de confuzie.